НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ННК «Інститут прикладного системного аналізу»
(повна назва інституту/факультету)

Кафедра системного проектування
(повна назва кафедри)

«На правах рукопису»
УДК _004.852

«До захисту допущено»

Завідувач кафедри
_____ _А.І. Петренко
(підпис) (ініціали, прізвище)

“___”_____2018 р.

Магістерська дисертація

зі спеціальності (спеціалізації)__122 – комп’ютерні науки та інформаційні технології (Системне проектування сервісів)_____
(код і назва спеціальності)

на тему: Методи аналізу та фільтрації потоків даних надвеликих обсягів. Колаборативна фільтрація на підставі аналізу змісту даних та фільтрація на підставі оцінки подібності груп даних та подібності користувачів даних_____

Виконав: студент _6_ курсу, групи ДА-61м
(шифр групи)

_____Прасолов Андрій Павлович _____ _____
(прізвище, ім’я, по батькові) (підпис)

Науковий керівник _____проф., д.т.н., Рогоза В.С._____ _____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Консультант Розроблення стартап-проекту _____ _____
(назва розділу) (науковий ступінь, вчене звання, , прізвище, ініціали) (підпис)

Рецензент _____ _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2018 року

Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»

Інститут/факультет ННК «Інститут прикладного системного аналізу»
(повна назва)

Кафедра Системного проектування
(повна назва)

Рівень вищої освіти – другий (магістерський) за освітньо-професійною (освітньо-науковою) програмою

зі спеціальності (спеціалізації) 122 – комп'ютерні науки та інформаційні технології (Інформаційні системи та технології проектування)
(код і назва)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ А.І. Петренко
(підпис) (ініціали, прізвище)
«___»_____2018р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Прасолову Андрію Павловичу
(прізвище, ім'я, по батькові)

1. Тема дисертації «Методи аналізу та фільтрації потоків даних надвеликих обсягів. Колаборативна фільтрація на підставі аналізу змісту даних та фільтрація на підставі оцінки подібності груп даних та подібності користувачів даних»
науковий керівник дисертації Рогоза В.С., д.т.н., проф.,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «10» *травня* 2018 р. №1028-с

2. Строк подання студентом дисертації _____

3. Об'єкт дослідження Колаборативна фільтрація

4. Предмет дослідження Алгоритми колаборативної фільтрації, які можна застосовувати в рекомендаційних системах

5. Перелік завдань, які потрібно розробити провести огляд існуючих методів та технологій побудови рекомендаційних систем на основі колаборативної фільтрації, розробити програму з реалізацією основних алгоритмів колаборативної

фільтрації та провести їх глибокий аналіз та порівняння, спробувати знайти способи вдосконалення існуючих методів реалізації, оформити роботу на основі отриманих результатів

6. Орієнтовний перелік публікацій Прасолов А.П.. Методи колаборативної фільтрації у рекомендаційних системах / А.П.Прасолов. // Міжнародний науковий журнал "Інтернаука". – 2018. – №8..

7. Консультанти розділів дисертації*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Реалізація стартап-проекту			

8. Дата видачі завдання 01.02.2018

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Строк виконання етапів магістерської дисертації	Примітка
1	Отримання завдання	01.02.2018	
2	Збір інформації та аналіз літератури	15.02.2018	
3	Проведення огляду існуючих методів колаборативної фільтрації	28.02.2018	
4	Реалізація алгоритмів колаборативної фільтрації та їх тестування.	13.04.2018	
5	Аналіз та порівняння результатів моделювання	25.04.2018	
6	Оформлення дипломної роботи	30.04.2018	
7	Отримання допуску до захисту та подача роботи в ДЕК	10.05.2018	

Студент _____ Прасолов А.П.
(підпис) (ініціали, прізвище)

Науковий керівник дисертації _____ Рогоза В.С.
(підпис) (ініціали, прізвище)

* Консультантом не може бути зазначено наукового керівника

РЕФЕРАТ НА МАГІСТЕРСЬКУ ДИСЕРТАЦІЮ

виконану на тему: Методи аналізу та фільтрації потоків даних надвеликих обсягів. Колаборативна фільтрація на підставі аналізу змісту даних та фільтрація на підставі оцінки подібності груп даних та подібності користувачів даних.

студентом: Прасоловим Андрієм Павловичем

Робота виконана на 78 сторінках, містить 22 ілюстрації, 21 таблиця. При підготовці використовувалась література з 20 джерел.

Актуальність теми

Людство не стоїть на місці. Постійно створюються нові продукти, знімаються нові фільми та створюється нова музика. Конкуренція тільки зростає і кожен намагається потрапити на ринок продемонструвавши свій продукт, його переваги та можливості. Це призводить до того, що на сьогодні ми маємо величезний вибір альтернатив. Йдучи на роботу слухаємо декілька з сотень тисяч пісень, ввечері дивимось один з великої кількості фільмів та купуємо певний товар з поміж безлічі інших.

Чим далі, тим важче людині робити вибір, тож програми допомагають скоротити кількість альтернатив, обравши для нас лише ті товари, що підходять саме для нас.

Алгоритми колаборативної фільтрації широко використовуються різними сайтами та інтернет-сервісами для різноманітних завдань, таких як рекомендації фільмів та музики, рекомендації нових друзів в соціальних мережах та навіть простий пошук інформації.

Мета та задачі дослідження

Метою роботи є дослідження та реалізація алгоритмів колаборативної фільтрації, їх аналіз та порівняння. Пошук ситуацій, у яких той чи інший алгоритм працює найкраще. Розгляд можливості комбінації декількох алгоритмів.

Рішення поставлених завдань та досягнуті результати

В ході виконання магістерської роботи було досліджено алгоритми колаборативної фільтрації, а саме Item-Based, User-Based, NMF, PMF, BPMF, Reg SVD та Slope One. Було створено програму, в якій було реалізовано перераховані алгоритми.

За допомогою створеної програми було проведено аналіз та порівняння алгоритмів. Було проведено аналіз точності алгоритмів, досліджено як алгоритми працюють з даними різного рівня розрідженості. Проведено дослідження швидкості роботи алгоритмів.

Об’єкт досліджень

Рекоментадійні системи.

Предмет досліджень

Алгоритми колаборативної фільтрації, які можна застосовувати в рекомендаційних системах.

Методи досліджень

Для вирішення проблеми в даній роботі використовуються методи аналізу і синтезу, системного аналізу, порівняння, логічного узагальнення результатів.

Наукова новизна

Наукова новизна роботи полягає у порівнянні великої кількості алгоритмів колаборативної фільтрації, які раніше не порівнювалися та у пошуці ситуацій, у яких той чи інший алгоритм проявляє себе найкращим чином.

Практичне значення одержаних результатів

Отримані результати можуть використовуватись у майбутніх дослідженнях за напрямком дослідження алгоритмів колаборативної фільтрації. Також, за допомогою результатів даної роботи можна обрати вдалий підхід до побудови оптимальної рекомендаційної системи.

Публікації

Прасолов А.П.. Методи колаборативної фільтрації у рекомендаційних системах / А.П.Прасолов. // Міжнародний науковий журнал "Інтернаука". – 2018. – №8.

Ключові слова

Колаборативна фільтрація, рекомендаційні системи, item-based, user-baser, NMF, PMF. BPMF, Regularized SVD, Slope One.

РЕФЕРАТ НА МАГИСТЕРСКУЮ ДИССЕРТАЦИЮ

Исполненную по теме: Методы анализа и фильтрации потоков данных сверхбольших объемов. Коллаборативная фильтрация на основе анализа содержания данных и фильтрация на основе оценки схожести групп данных и схожести пользователей данных

студентом: Прасоловым Андреем Павловичем

Робота исполнена на 78 страницах, содержит 22 иллюстрации, 21 таблицы. При подготовке использовалась литература из 20 источников.

Актуальность темы

Человечество не стоит на месте. Постоянно создаются новые продукты, снимаются новые фильмы и создается новая музыка. Конкуренция только возрастает и каждый пытается попасть на рынок показав свой продукт, его преимущества и возможности. Это приводит к тому, что на сегодня мы имеем огромный выбор альтернатив. Идя на работу слушаем несколько из сотен тысяч песен, вечером смотрим один из огромного множества фильмов и покупаем определенный товар из множества других.

Чем дальше, тем труднее человеку делать выбор, поэтому программы помогают сократить количество альтернатив, выбрав для нас только те товары, которые подходят именно нам.

Алгоритмы коллаборативной фильтрации широко используются разными сайтами и интернет-сервисами для разнообразных задач, таких как рекомендации фильмов и музыки, рекомендации новых друзей в социальных сетях и даже простой поиск информации.

Цели и задачи исследования

Целью работи является исследование и реализация алгоритмов коллаборативной фильтрации, их анализ и сравнение. Поиск ситуаций, в которых

тот или иной алгоритм работает наилучше. Рассмотрение возможности комбинации нескольких алгоритмов.

Решения поставленных задач и полученные результаты

В ходе выполнения магистерской работы было исследовано алгоритмы коллаборативной фильтрации, а именно Item-Based, User-Based, NMF, PMF, BPMF, Reg SVD и Slope One. Было создано программу, в которой было реализовано перечисленные алгоритмы.

При помощи созданной программы было призведено анализ и сравнение алгоритмов. Было проведено анализ точности алгоритмов, изучено как алгоритмы работают с даннымы разного уровня разреженности. Произведено исследование быстродействия работы алгоритмов.

Объект исследования

Коллаборативная фильтрация.

Предмет исследования

Алгоритмы коллаборативной фильтрации, которые можна применить в рекомендационных системах.

Методы исследований

Для решения проблемы в данной работе использовались методы анализа и синтеза, системного анализа, сравнения, логического обобщения результатов.

Научная новизна

Научная новизна работы состоит в сравнении большого количества алгоритмов коллаборативной фильтрации, которые раньше не сравнивались и в поиске ситуацый, в которых той или иной алгоритм показывает себя лучшим образом.

Практическое значение полученных результатов

Полученные результаты могут быть использованы в будущих исследованиях в направлении исследований алгоритмов коллаборативной

фильтрации. Также, при помощи результатов этой работы можна выбрать удачный подход для построения оптимальной рекомендационной системы.

Публикации

Прасолов А.П.. Методы коллаборативной фильтрации в рекомендационных системах / А.П.Прасолов. // Международный научный журнал "Интернаука". – 2018. – №8.

Ключевые слова

Коллаборативна фильтрация, рекомендационные системы, item-based, user-baser, NMF, PMF. BPMF, Regularized SVD, Slope One.

ABSTRACT ON MASTER'S THESIS

on topic: Methods of analysis and filtering of streams of data extra-large size. Collaborative filtering based on data analysis and filtering based on the assessment of the similarity of data groups and the similarity of data users

student: Andrii Prasolov

Work carried out on 78 pages containing 22 figures, 21 tables. The paper was written with references to 20 different sources.

Topicality

Humanity does not stand still. Constantly creating new products, filming new movies and creating new music. Competition only extends and everyone tries to enter the market by demonstrating their product, its benefits, and opportunities. This leads to the fact that today we have a large choice of alternatives. When going to work, we listen one of the thousands of songs, in the evening we watch one of a large number of films and buy a certain item from among many others.

So far, harder and harder a person to make a choice, so the programs help reduce the number of alternatives for us choosing only those products that are suitable for us.

Collaborative filtering algorithms are widely used by various sites and Internet services for a variety of tasks such as movie and music recommendations, new friends' recommendations in social networking and even simple information searches.

Purpose

The purpose of the work is to research and implement collaborative filtration algorithms, their analysis and comparison. Searching for situations in one or another algorithm works better. Considering the combination of several algorithms.

Solution

In the course of the master's thesis collaborative filtration algorithms, namely Item-Based, User-Based, NMF, PMF, BPMF, Reg SVD and Slope One were investigated. A program was created in which the listed algorithms were implemented.

The created program analyzed and compared algorithms. An analysis of the accuracy of the algorithms was performed, and the algorithms were investigated based on the data of different levels of density. The study of the performance of the algorithms is carried out.

Object of research

Collaborative filtration.

Subject of research

Collaborative filtering algorithms that can be used in recommendation systems.

Research methods

To solve the problem in this paper were used methods of analysis and synthesis, system analysis, comparison, logical generalization of results.

Scientific novelty

The scientific novelty of the work is to compare a large number of collaborative filtration algorithms and to find situations in which one or another algorithm manifests itself in the best way.

The practical value of the results

The obtained results can be used in future research in the direction of research of collaborative filtration algorithms. Also, with the help of the results of this work you can choose a successful approach to building an optimal recommendation system.

Publications

Prasolov A.P. Methods of collaborative filtering in recommender systems / Prasolov A.P. // International scientific magazine "Internet Science". - 2018 - №8.

Keywords

Collaborative filtering, recomendation systems, item-based, user-baser, NMF, PMF. BPMF, Regularized SVD, Slope One.

ЗМІСТ

Перелік умовних позначень

РС – Рекомендаційна Система

КФ – Колаборативна Фільтрація

CF – Collaborative Filtering

SVD – Singular Value Decomposition

RSVD – Regularized Singular Value Decomposition

NMF – Non-negative Matrix Factorization

PMF – Probabilistic matrix factorization

BPMF – Bayesian Probabilistic matrix factorization

ВСТУП

Життя людства розділилося на два етапи: до створення мережі Інтернет, та після. Велику частку свого життя люди проводять, проживають в інтернеті. Вони там знайомляться, спілкуються, розважаються, навчаються та купують різні товари та послуги.

Звернемо увагу на продаж товарів. Як це відбувається у реальному світі? Ми приходимо у магазин і продавець/консультант, в міру своїх навичок, намагається продати нам якомога більше товарів та хоче зацікавити вас повернутися до магазину ще. Як це відбувається? Він намагається підібрати товар/послугу, яка якомога краще підходить саме вам.

Інтернет магазини та різноманітні веб-сервіси вирішили не відмовлятись від настільки гарного підходу збільшити свої прибутки або популярність ресурсу, що в свою чергу також збільшить прибутки. Тож вони (інтернет магазини, торгівельні платформи, всілякі веб-сервіси) застосовують різного рівня якості рекомендаційні системи. Також, ще однією з причин звернення уваги на рекомендаційні системи є те, що в інтернеті існує величезна кількість всіляких товарів та послуг і людині просто фізично важко зробити вибір з такого різноманіття. Тут рекомендаційні системи і приходять на допомогу, адже значно простіше зробити вибір з декількох одиниць, які для тебе підготувала система, ніж переглядати тисячі і тисячі товарів, які абсолютно вас не цікавлять.

Рекомендаційні системи не можна назвати новою технологією, адже перші їх версії були реалізовані ще два десятки років тому, але активного розвитку та розповсюдження вони зазнали лише в останні роки.

Найбільш успішним типом рекомендаційних систем є системи, побудовані на основі колаборативної фільтрації. Це такий тип систем, які будують рекомендації та шукають потенційно цікаві для користувача товари/послуги

базуючись на попередніх покупках/діях даного користувача та користувачів з схожими інтересами.

Ціллю даної роботи є реалізація, глибокий аналіз і порівняння найбільш розповсюджених та найбільш продуктивних алгоритмів колоборативної фільтрації.

Ми розглянемо існуючі алгоритми. Розглянемо їх переваги те недоліки. Проведемо експеременти їх роботи на різних наборах даних з різним рівнем розрідженності. Визначимо які з них краще працюють на більш розріджених ноборах даних, а які навпаки. Розглянемо системи виміру точності надання рекомендацій. За результатами експерементів визначимо які алгоритми краще працюють в режимі реального часу, а які можуть дати кращі рекомендації, незважаючи на те, що для цього знадобиться значно більше часу на обчислення.

Також слід пам'ятати, що колаборативна фільтрація, а точніше, рекомендаційні системи на її основі, можуть використовуватися у багатьох галузях. Не потрібно обмежувати цей клас програмного забезпечення лише рекомендаціями фільмів чи товарів, адже це частина нашого майбутнього, яка неодмінно вплине не тільки на життя наших нащадків, а й на наше, адже колесо прогресу та еволюції вже запущено і воно лише набирає швидкості.

Тож, можна сміливо зазначити, що колоборативна фільтрація є дуже актуальною темою дослідження.

1 РЕКОМЕНДАЦІЙНІ СИСТЕМИ

1.1 Визначення

Рекомендаційна система — підклас системи фільтрації інформації, яка будує рейтинговий перелік об'єктів (фільми, музика, книги, новини, веб-сайти), яким користувач може надати перевагу. Для цього використовується інформація з профілю користувача [1].

Рекомендаційні системи зберігають дані про користувачів, створюючи їх профілі і записуючи туди їх вподобання. Більшість рекомендаційних систем в профілях користувачів зберігають набори оцінок, оцінки можуть мати різний вигляд: «подобається» і «не подобається», значення в якомусь діапазоні (1-5), чи у якийсь інший спосіб. Чим вищу оцінку користувач ставить товару, тим більше цей товар сподобався йому (користувачу).

Потім товари зіставляються з результатами попереднього досвіду користувача (з товарами, які він вже оцінив). Також різні товари можуть мати різну вагу в цьому процесі, в залежності від розміру оцінки, яку виставив їм користувач. Найбільш відповідні товари, тобто ті, що при зіставленні отримали найбільший бал/найбільшу вагу, рекомендуються користувачу.

1.2 Конкурс «Netflix Prize» від компанії Netflix на розробку кращого алгоритму КФ

Рекомендаційні системи з'явилися в інтернеті досить давно, близько 20-25 років тому. Однак справжній підйом в цій області трапився приблизно 12 років тому, коли відбулося змагання Netflix Prize.

Компанія Netflix тоді давала в прокат не цифрові копії, а розсилала VHS-касети і DVD.

Для них було дуже важливо підвищити якість рекомендацій. Чим краще Netflix рекомендує своїм користувачам фільми, тим більше фільмів вони беруть в прокат. Відповідно, зростає і прибуток компанії.

У 2006 році вони запустили змагання Netflix Prize. Вони виклали у відкритий доступ зібрані дані: близько 100 мільйонів оцінок за п'ятибальною шкалою з зазначенням ID проставити їх користувачів. Учасники змагання повинні були якомога краще передбачати, яку оцінку поставить певному фільму той чи інший користувач. Якість передбачення вимірювалося за допомогою метрики RMSE (середньо-квадратичне відхилення).

У Netflix вже був алгоритм, який передбачав оцінку з якістю 0.9514 за метрикою RMSE. Завдання було поліпшити прогноз хоча б на 10% - до 0.8563. Переможцю був обіцяний приз в $ 1 000 000. Змагання тривало приблизно три роки.

За перший рік якість поліпшили на 7%, далі все трохи сповільнилося. Але в кінці дві команди з різницею в 20 хвилин надіслали свої рішення, кожне з яких проходило поріг в 10%, якість у них була однакова з точністю до четвертого знака. У задачі, над якою безліч команд билося три роки, все вирішили якихось двадцять хвилин. Команда, що запізнилась (як і багато інших, які брали участь в конкурсі),

залишилися ні з чим, однак сам конкурс дуже сильно прискорив розвиток в цій галузі [2].

1.3 Підходи та принципи

Основне припущення колаборативної фільтрації полягає в наступному: ті, хто однаково оцінювали будь-які предмети в минулому, схильні давати схожі оцінки інших предметів і в майбутньому.

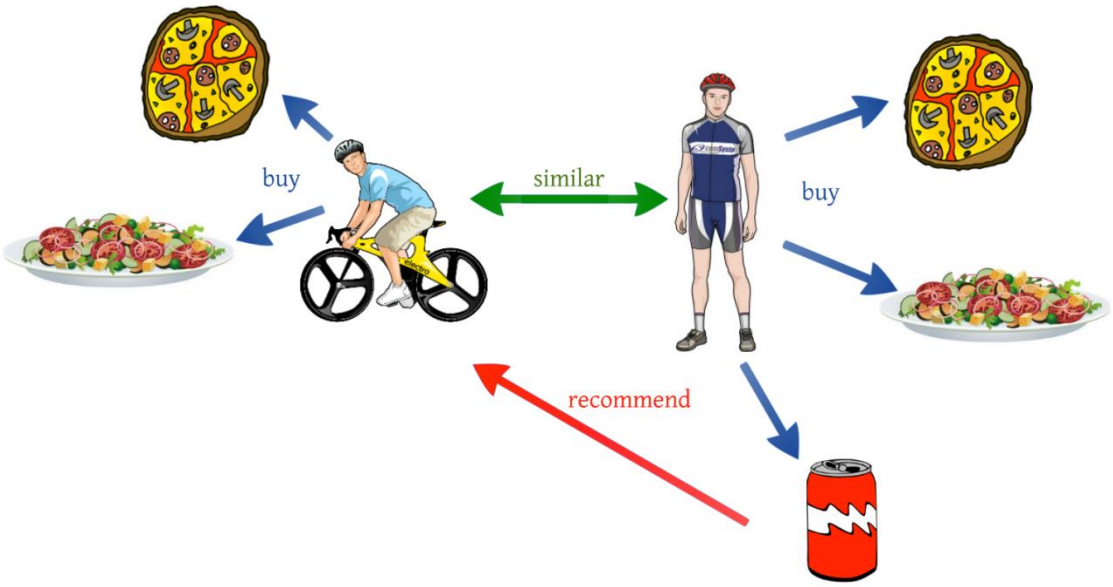

Рисунок 1.1 – Принцип роботи рекомендаційних систем

Підходи в побудові рекомендаційних систем [3]:

- Підхід на підставі ознакових описів (contant-based): передбачає, що про користувачів та товари/об’єкти відомо досить багато інформації, за допомогою якої можна будувати рекомендації.
- Колаборативна фільтрація: рекомендації будуються на підставі взаємодії користувачів з таварами/об’єктами.
- Гібридний підхід: використовує композицію двох попередніх груп алгоритмів.

Рисунок 1.2 – Підходи в побудові рекомендаційних систем [3]

1.4 Вимір якості рекомендацій

Якщо ми хочемо поліпшити якість рекомендацій, нам потрібно навчитися його вимірювати. Для цього алгоритм, навчений на одній вибірці - навчальній, перевіряється на іншій - тестовій. Netflix запропонував вимірювати якість рекомендацій по метриці RMSE[14]:

$$RMSE = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} (\hat{r}_{ui} - r_{ui})^2} \tag{1.1}$$

Також існує ще одна метрика – mean absolute error (MAE), яку ми будемо використовувати надалі.

$$MAE = \frac{1}{n}\sum_{u,i}|p_{u,i} - r_{u,i}| \tag{1.2}$$

MAE - це середня вертикальна відстань між кожною точкою та лінією Y = X, яка також відома як "one-to-one" лінія. MAE також є середньою горизонтальною відстанню між кожною точкою і лінією Y = X[13].

Також ще слід зазначити ще деякі властивості рекомендацій: на сприйняття рекомендацій впливає не тільки якість ранжирування, а й деякі інші характеристики. Серед них, наприклад, різноманітність (не варто видавати користувачу фільми тільки на одну тему або з однієї серії), несподіванка (якщо рекомендувати дуже популярні фільми, то такі рекомендації будуть занадто банальними і майже марними), новизна (багатьом подобаються класичні фільми, але рекомендаціями зазвичай користуються, щоб відкрити для себе щось нове) і багато іншого.

1.5 Висновки

В цьому розділі було розглянуто рекомендаційні системи. Було надано визначення рекомендаційної системи, наведено декілька фактів з історії створення та розвитку рекомендаційних систем. Коротко описано основні підходи та принципи побудови таких систем. Також було описано системи виміру якості надання рекомендацій і вплив латентних (прихованих) параметрів на точність рекомендації. Цей розділ було написано для надання базових понять, які необхідня для розуміння подальших частин роботи.

2 КОЛАБОРАТИВНА ФІЛЬТРАЦІЯ

2.1 Визначення

Колаборативна фільтрація, спільна **фільтрація** — метод, який використовується деякими рекомендаційними системами. Колаборативна фільтрація має два значення: вузьке і більш загальне. В цілому, колаборативна фільтрація — процес фільтрації інформації або зразків за допомогою методів за участю співробітництва між декількома агентами, точками зору, джерелами даних і т. д. Застосування колаборативної фільтрації, як правило, пов'язане з дуже великими наборами даних. Колаборативні методи фільтрації були застосовані до різних видів даних, зокрема до таких як зондування та моніторинг даних, які виникають при розвідці корисних копалин на великих площах; до фінансових даних, таких як установи фінансових послуг, які об'єднують багато фінансових джерел; або в електронній торгівлі та веб-додатках, що зосереджуються на даних користувача, і т. д. Решта цієї дискусії зосереджена на колаборативній фільтрації даних, призначених для користувача, хоча деякі з методів та підходів можуть застосовуватися так само і у багатьох інших випадках[3].

У більш новому, вужчому значенні колаборативна фільтрація — це один з методів побудови прогнозу в рекомендаційних системах, який використовує відомі переваги (оцінки) групи користувачів для прогнозування невідомих переваг іншого користувача.[1] Основне припущення колаборативної фільтрації полягає в наступному: ті, хто однаково оцінювали будь-які предмети в минулому, схильні давати схожі оцінки інших предметів і в майбутньому. Наприклад, за допомогою колаборативної фільтрації музичний додаток здатний прогнозувати, яка музика сподобається користувачеві, маючи неповний список його переваг (симпатій та

антипатій).[2]Прогнози складаються індивідуально для кожного користувача, хоча інформація, що використовується, зібрана від багатьох учасників. Це відрізняє колаборативну фільтрацію від більш простого підходу, дає усереднену оцінку для кожного об'єкта інтересу, наприклад того, що базується на кількості поданих за нього голосів. Дослідження в даній області активно ведуться і в наш час, що зокрема обумовлюється наявністю невирішених проблем у методі колаборатівної фільтрації[3].

2.2 Проблеми колаборативної фільтрації

Розрідженість даних.

Рекомендаційні систами, зазвичай, у своїй основі, базуються на тому, що вони включають в себе величезну кількість користувачів та товарів. Але користувачі не завжди оцінюють товари, навіть якщо придбали їх. В результаті данні в матриці «користувач-предмет» виходять дуже розріджені. Ця проблема особливо гостро стоїть для нещодавно створених рекомендаційних систем.

Також ця проблема (розрідженість даних) загострює проблему холодного старту[3].

Масштабованість.

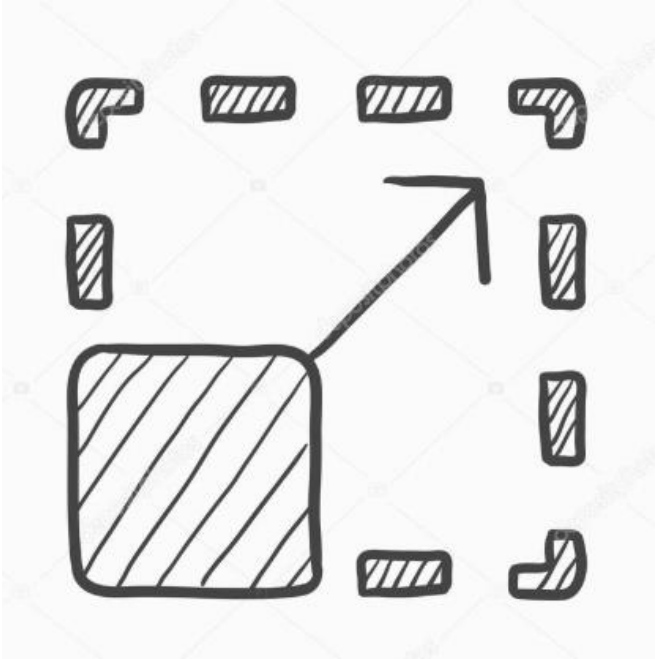

Рисунок 2.1 - Масштабованість

Зі збільшенням кількості користувачів в системі, з'являється проблема масштабованості. Наприклад, маючи 10 мільйонів покупців *O(M)* і мільйон предметів *O(N),* алгоритм колаборативної фільтрації зі складністю рівній *O(NM)* вже занадто складний для розрахунків. Також, багато систем повинні моментально реагувати на онлайн запити від всіх користувачів, незалежно від історії їх покупок і оцінок, що вимагає ще більшої масштабованості[3].

Проблема холодного старту.

Нові предмети або користувачі представляють велику проблему для рекомендаційних систем. Частково проблему допомагає вирішити підхід, заснований на аналізі вмісту, так як він покладається не на оцінки, а на атрибути, що допомагає включати нові предмети в рекомендації для користувачів. Однак проблему з наданням рекомендації для нового користувача вирішити складніше[3].

Синонімія.

Синонімією називається тенденція схожих і однакових предметів мати різні імена. Більшість рекомендаційних систем не здатні виявити ці приховані зв'язки і тому відносяться до цих предметів як до різних. Наприклад, «фільми для дітей» та «дитячий фільм» відносяться до одного жанру, але система сприймає їх як різні[3].

Шахрайство.

Рисунок 2.2 – Шахрайство

Рекомендаційні системи – це таке місце де люди можуть ставити будь-які оцінки будь-яким товарам. Вони можуть давати лише позитивні оцінки своїм товарам (хоча їх якість може бути далекою від ідеалу) і погані, товарам конкурентів. Також РС можуть сильно вплинути на прибутки компаній через своє широке розповсюдження в інтернет комерції.. Отже, недобросовісні виробники намагаються шахрайським чином підняти рейтинг власних продуктів та понизити рейтинги своїх конкурентів[3].

Різноманітність.

Колаборативна фільтрація спочатку має збільшити різноманітність, щоб дозволяти відкривати користувачам нові продукти з незліченної множини. Однак деякі алгоритми, зокрема основні на продажах і рейтингах, створюють дуже

складні умови для просування нових і маловідомих продуктів, так як їх заміщають популярні продукти, які давно перебувають на ринку. Це в свою чергу тільки збільшує ефект «багаті стають ще багатшими» і приводить до меншої різноманітності[3].

Білі ворони.

«Білі ворони» - це такий тип користувачів, думка яких майже завжди не збігається з думкою більшості. В них, можна сказати, унікальний смак, тому їм неможливо щось порекомендувати. Однак кількість таких людей невелика і в реальному житті така ситуація теж їх не покидає, тож дослідження з виправлення цієї проблеми не ведуться [3].

2.3 Алгоритми колаборативної фільтрації

2.3.1 Алгоритм колаборативної фільтрації Item-Base

Метод КФ, який базується на тому, що користувачу сподобаються товари, схожі на ті, які він вже обирав.

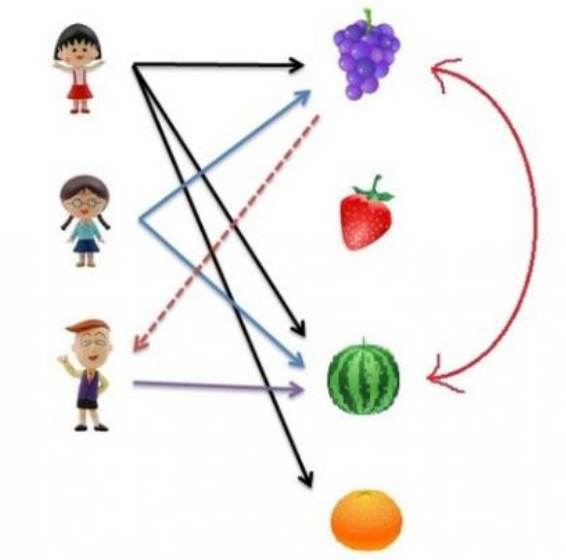

Рисунок 2.3 – Item-Base алгоритм [4]

Користувач A характеризується об'єктами, які він переглянув або оцінив. Для кожного об'єкта з обраних/оцінених визначається m об'єктів-сусідів, тобто знаходяться m найбільш схожих об'єктів з точки зору переглядів / оцінок користувачів. При побудові РС для фільмів, m приймає значення від 10 до 30. Всі об'єкти-сусіди об'єднуються в колекцію, з якої виключаються об'єкти, переглянуті або оцінені користувачем A. A з останків колекції будується топ n рекомендацій. Таким чином, при item-based підході у створенні рекомендацій беруть участь всі користувачі, яким сподобався той чи інший об'єкт з колекції.

$$\hat{r}_{ui} = \bar{r}_i + \frac{1}{\sum_{i' \in R(u)} |sim(i,i')|} \sum_{i' \in R(u)} sim(i,i')(r_{u,i'} - \bar{r}_{i'}) \quad (2.1)$$

2.3.2 Алгоритм колаборативної фільтрації UserBase

Метод КФ, який базується на тому, що користувачу сподобаються товари, які були обрані користувачами, схожими на нього.

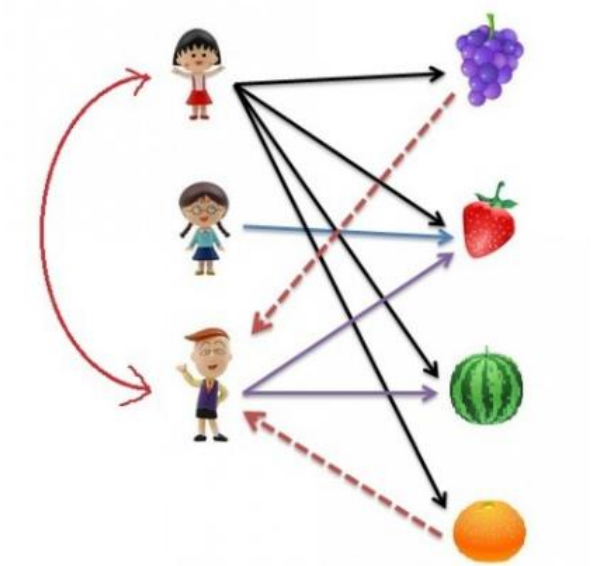

Рисунок 2.4 – User-Base алгоритм [4]

$$\hat{r}_{ui} = \bar{r}_i + \frac{1}{\sum_{i' \in R(u)} |sim(i,i')|} \sum_{i' \in R(u)} sim(i,i')(r_{u,i'} - \bar{r}_{i'}) \quad (2.2)$$

Міра схожості $sim(u,u')$ обчислюється за матрицею оцінок *R*. Найбільш загальновживана метрика схожості – кореляція Пірсона і косинусна відстань рядків (стовпців) матриці[5].

$$\mathbf{sim}(u,u') = \frac{\sum_{i\in \mathbf{R}(u)\cap \mathbf{R}(u')}\left(r_{u,i}-\bar{r}_u\right)\left(r_{u',i}-\bar{r}_{u'}\right)}{\sqrt{\sum_{i\in \mathbf{R}(u)\cap \mathbf{R}(u')}\left(r_{u,i}-\bar{r}_u\right)^2 \sum_{i\in \mathbf{R}(u)\cap \mathbf{R}(u')}\left(r_{u',i}-\bar{r}_{u'}\right)^2}} \tag{2.3}$$

$$\mathbf{sim}(u,u') = \frac{\sum_{i\in \mathbf{R}(u)\cap \mathbf{R}(u')} r_{u,i}\, r_{u',i}}{\sqrt{\sum_{i\in \mathbf{R}(u)} r_{u,i}^2 \sum_{i\in \mathbf{R}(u')} r_{u',i}^2}} \tag{2.4}$$

2.3.3 Алгоритм колаборативної фільтрації Probabilistic matrix factorization (PMF)

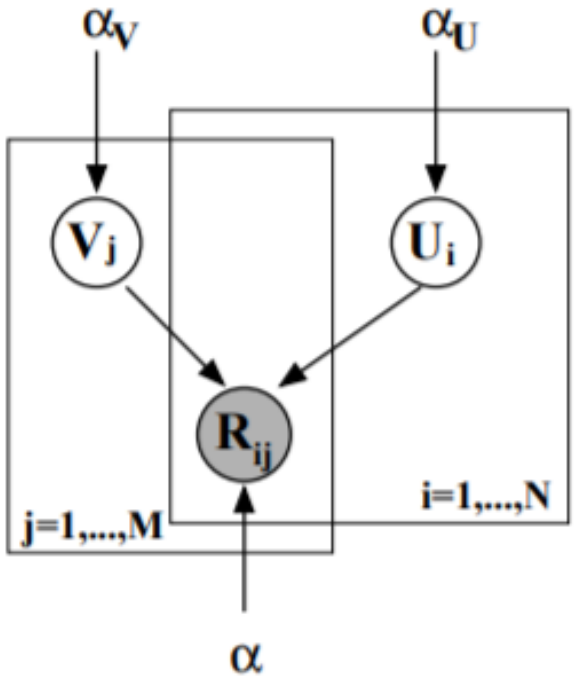

Рисунок 2.5 – PMF модель[6]

Ймовірнісна матрична факторизація (PMF) - імовірнісна лінійна модель з гаусовим спостережним шумом. Припустимо, у нас є N користувачів і M фільми. Нехай Rij стане рейтинговим значенням для користувача i для фільму j, Ui та Vj представляють D-розмірні специфічні для користувача і для фільму вектори

латентної функції, відповідно[6]. Умовний розподіл за спостереженими оцінками $R \in \mathbb{R}^{N \times M}$ (термін правдоподібності) та попередні розподіли за $U \in \mathbb{R}^{D \times N}$і $V \in \mathbb{R}^{D \times \bar{M}}$ визначаються за допомогою:

$$p(R|U,V,\alpha) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(R_{ij}|U_i^T V_j, \alpha^{-1})\right]^{I_{ij}} \tag{2.5}$$

$$p(U|\alpha_U) = \prod_{i=1}^{N}\mathcal{N}(U_i|0,\alpha_U^{-1}I) \tag{2.6}$$

$$p(V|\alpha_V) = \prod_{j=1}^{M}\mathcal{N}(V_j|0,\alpha_V^{-1}I), \tag{2.7}$$

Де $\mathcal{N}(x|\mu,\alpha^{-1})$ позначає гауссовський розподіл з середнім μ і точністю α, а Iij - індикаторна змінна, яка дорівнює 1, якщо користувач оцінив фільм j і рівний 0 в протилежному випадку. Навчання в цій моделі виконується шляхом максимізації лог-попереду над фільмом та функціями користувача з фіксованими гіперпараметрами (тобто дисперсією шумів спостережень та попередніми відхиленнями):

$$\begin{aligned}\ln p(U,V|R,\alpha,\alpha_V,\alpha_U) &= \ln p(R|U,V,\alpha) + \\ &+ \ln p(U|\alpha_U) + \ln p(V|\alpha_V) + C,\end{aligned} \tag{2.8}$$

де C - це константа, яка не залежить від параметрів. Максимізація цього поперечного розподілу по відношенню до U та V еквівалентно мінімізації функції помилок суми квадратів з членами квадратичної регуляризації

$$\begin{aligned}E &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\left(R_{ij} - U_i^T V_j\right)^2 \\ &+ \frac{\lambda_U}{2}\sum_{i=1}^{N} \parallel U_i \parallel_{\mathrm{Fro}}^2 + \frac{\lambda_V}{2}\sum_{j=1}^{M} \parallel V_j \parallel_{\mathrm{Fro}}^2\end{aligned} \tag{2.9}$$

де λU = αU / α, λV = αV / α, і $\|\cdot\|_{\mathrm{Fro}}^2$ позначає норму Фробеніуса. Локальний мінімум об'єктивної функції, заданої у рівнянні[6].

2.3.4 Алгоритм колаборативної фільтрації Bayesian probabilistic matrix factorization (BPMF)

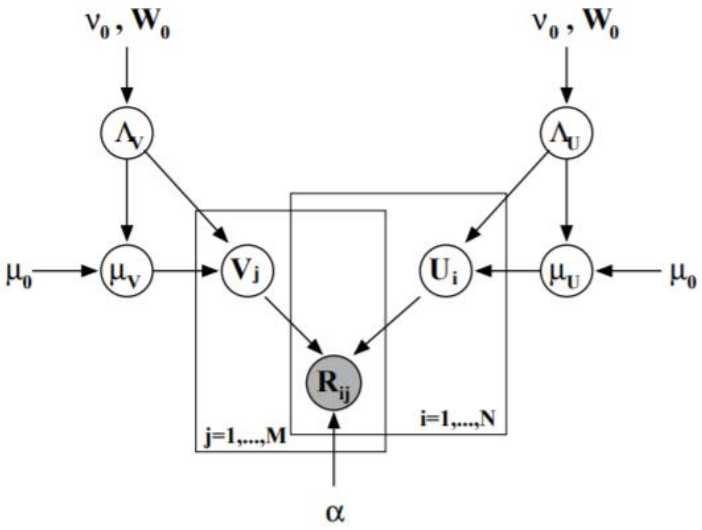

Рисунок 2.6 – BPMF модель[6]

Попередні розподіли над векторами функцій користувача та фільму вважаються гауссовськими:

$$p(U|\mu_U, \Lambda_U) = \prod^{N} \mathcal{N}(U_i|\mu_U, \Lambda_U^{-1}) \tag{2.10}$$

$$p(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \mathcal{N}(V_i|\mu_V, \Lambda_V^{-1}) \tag{2.11}$$

Надалі ми розміщуємо прихильників Gaussian-Wishart на користувацькому та гіперпараметрах фільму ΘU = {μU, ΛU} і ΘV = {μV, ΛV}:

$$\begin{aligned} p(\Theta_U|\Theta_0) &= p(\mu_U|\Lambda_U)p(\Lambda_U) \\ &= \mathcal{N}(\mu_U|\mu_0, (\beta_0\Lambda_U)^{-1})\mathcal{W}(\Lambda_U|W_0, \nu_0) \end{aligned} \tag{2.12}$$

$$\begin{aligned} p(\Theta_V|\Theta_0) &= p(\mu_V|\Lambda_V)p(\Lambda_V) \\ &= \mathcal{N}(\mu_V|\mu_0, (\beta_0\Lambda_V)^{-1})\mathcal{W}(\Lambda_V|W_0, \nu_0) \end{aligned} \tag{2.13}$$

Тут W - розподіл Wishart з ν0 ступенями свободи і матриця масштабу D × D W0:

$$\mathcal{W}(\Lambda|W_0,\nu_0) = \frac{1}{C}|\Lambda|^{(\nu_0-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(W_0^{-1}\Lambda)\right) \quad (2.14)$$

де C - нормалізаційна константа. Для зручності ми також визначаємо Θ0 = {μ0, ν0, W0}. У наших експериментах ми також встановили ν0 = D і W0 на матрицю тотожності як для гіперпараметрів користувача, так і для фільмів, і вибираємо μ0 = 0 за симетрією[6].

2.3.5 Алгоритм колаборативної фільтрації Singular value decomposition (SVD)

Майже всі алгоритми колаборативної фільтрації мають такі недоліки, як холодний старт, тривіальність результатів рекомендацій тощо. Одним з досить нових алгоритмів, який знижує вплив типових проблем колаборативної фільтрації, виявився SVD алгоритм, який було створено саме для покращення результатів звичайних алгоритмів[7].

SVD - це методика факторизації матриць, яка зазвичай використовується для зменшення кількості функцій набору даних шляхом зменшення розмірів простору від N до K, де K < N. Однак для цілей систем рекомендацій нас цікавить лише матрична факторизація, де частини зберігають таку ж розмірність. Матрична факторизація виконується на матриці рейтингів користувацьких позицій. Матрична факторизація може розглядатися як пошук 2 матриць, продукт яких є вихідною матрицею.

$$expected\ rating = \hat{r}_{ui} = q_i^T p_u \quad (2.15)$$

`qi` і` pu` можна знайти таким чином, що різниця квадратних помилок між їх точками продукту та відомим рейтингом у матриці користувацького елемента є мінімальною.

$$minimum(p,q) \sum_{(u,i)\epsilon K} \left(r_{ui} - q_i^T . p_u\right)^2 \quad (2.16)$$

2.3.6 Алгоритм колаборативної фільтрації Regularized single value decomposition (RSVD)

Щоб наша модель могла добре узагальнювати і не перевищувати навчальний набір, ми запроваджуємо строк покарання для нашого рівняння мінімізації. Це представлено коефіцієнтом регуляризації λ, помноженої на квадратну суму величин векторів користувача та елементів[7].

$$minimum(p,q) \sum_{(u,i)\epsilon K} \left(r_{ui} - q_i^T . p_u\right)^2 + \lambda(\|q_i\|^2 + \|p_u\|^2) \quad (2.17)$$

2.3.7 Алгоритм колаборативної фільтрації Slope One

Slope One - це один з найпростіших підходів до рекомендацій на основі колаборативної фільтрації за подібністю предметів, але в той же час точність рекомендацій алгоритму порівняна з більш складними та ресурсоємними алгоритмами. Він був розроблений Даніелем Лемайром та Анною Маклахман в 2004 році і опублікований у 2005 році в статті.

У алгоритмі SlopeOne враховуються як інформація від інших користувачів, які оцінили той самий елемент, так і з інших предметів, оцінених одним і тим самим користувачем. Проте, схеми також розраховуються на точках даних, які не потрапляють ні в масив користувача, ні в масив об'єкта, але, однак, є важливою інформацією для прогнозування рейтингу[8].

Враховуючи набір тренувальних наборів χ та будь-які два пункти j та i з рейтингом uj та ui відповідно в деякій оцінці користувача u, ми розглянемо середнє відхилення пункту i стосовно пункту j як:

$$\mathrm{dev}_{j,i} = \sum_{u \in S_{j,i}(\chi)} \frac{u_j - u_i}{card(S_{j,i}(\chi))} \tag{2.18}$$

Будь-яка оцінка користувача, яка не містить як uj, так і ui, не входить до суми. Симметрична матриця, визначена devj,i може бути одночасно обчислена і швидко оновлюватися при введенні нових даних.

Враховуючи те, що devj, i + ui є прогнозом для uj даного ui, розумним предиктором може бути середнє значення всіх таких прогнозів.

$$P(u)_j = \frac{1}{card(R_j)} \sum_{i \in R_i} (\mathrm{dev}_{j,i} + u_i) \tag{2.19}$$

Можемо спростити формулу до:

$$P^{S1}(u)_j = \bar{u} + \frac{1}{card(R_j)} \sum_{i \in R_i} \mathrm{dev}_{j,i}. \tag{2.20}$$

2.3.8 Алгоритм колаборативної фільтрації Non negative matrix factorization (NMF)

Факторизация невід'ємних матриць (NMF) - це уявлення матриці V у вигляді добутку матриць W і H, в якому всі елементи трьох матриць невід'ємні[9].

Нехай таблиця V має розмір m x n. Позначимо через r ранг матриць W і H, як правило r << min (n, m). На відміну від точного уявлення матриці в SVD, в NMF маємо тільки наближену рівність.

$$V \approx WH. \tag{2.21}$$

Матриці W і H вибираються таким чином, щоб мінімізувати функцію втрати: D (V, WH) -> min. У нашому випадку D задається на основі дивергенції Кульбака-Лейблера

$$D(A, B) = \sum_{i,j} a_{ij}\ log(\frac{a_{ij}}{b_{ij}}) - a_{ij} + b_{ij}. \tag{2.22}$$

2.4 Висновки

В цьому розділі було досліджено алгоритми колаборативної фільтрації. Було надано визначення колаборативної фільтрації, наведено перелік недоліків, які зазвичай зустрічаються в алгоритмах колаборативної фільтрації.

Було перераховано алгоритми які буде використано в наступному розділі магістерської дисертації. До кожного алгоритму було надано короткий опис та формулу/формули, що демонструють принципи їх роботи.

Отже, цей розділ є гарним підґрунтям для подальшого аналізу алгоритмів колаборативної фільтрації.

3 РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Під час написання даної роботи було розроблено програмний продукт, що включає в себе реалізацію набору алгоритмів колаборативної фільтрації.

В цьому розділі буде розглянуто результати роботи різних алгоритмів колаборативної фільтрації за різних умов, таких як кількість користувачів, товарів (фільмів) та розрідженості матриці. Буде проведено аналіз цих алгоритмів, їх недоліків та переваг.

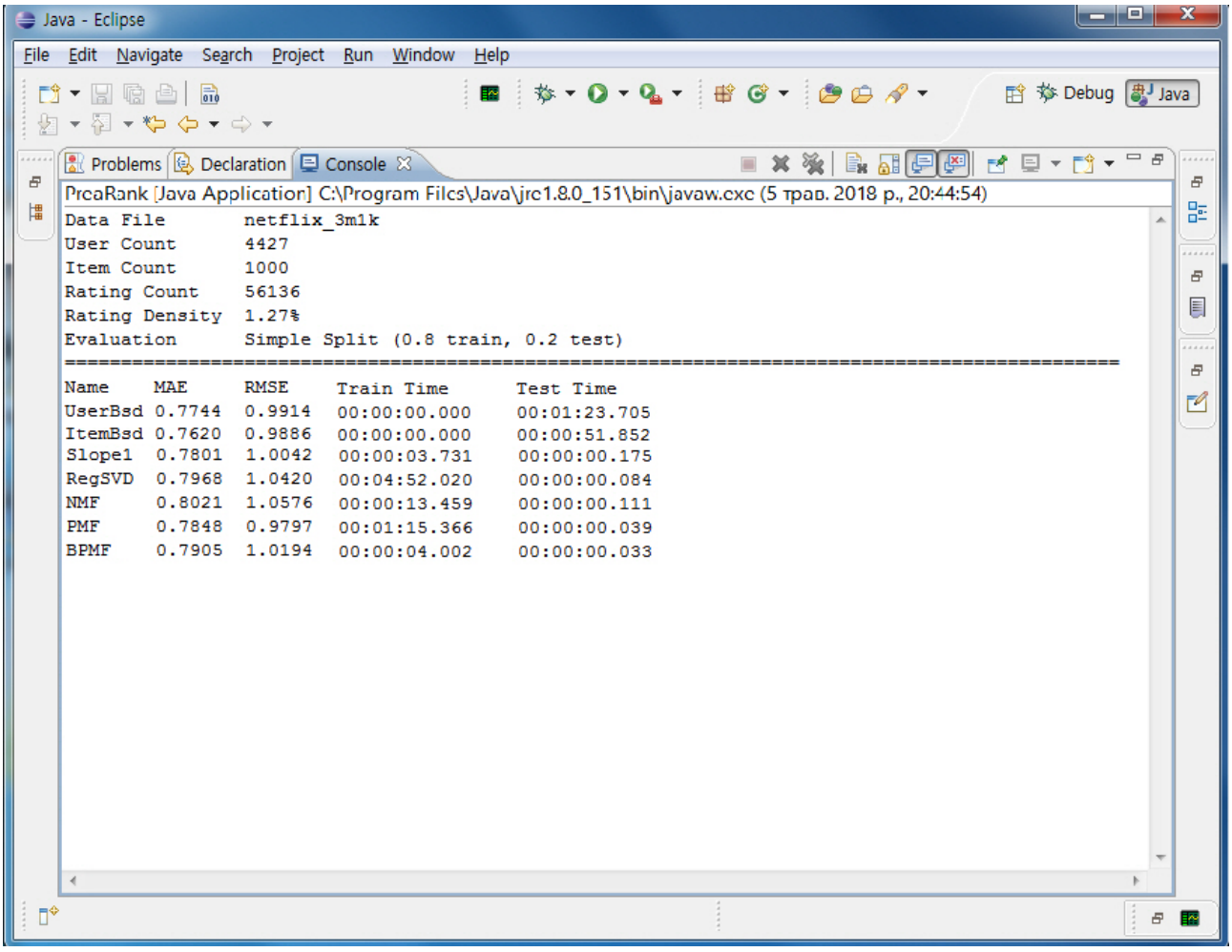

Рисунок 3.1 – скріншот виводу результатів роботи програми

3.1 Дослідження точності алгоритмів

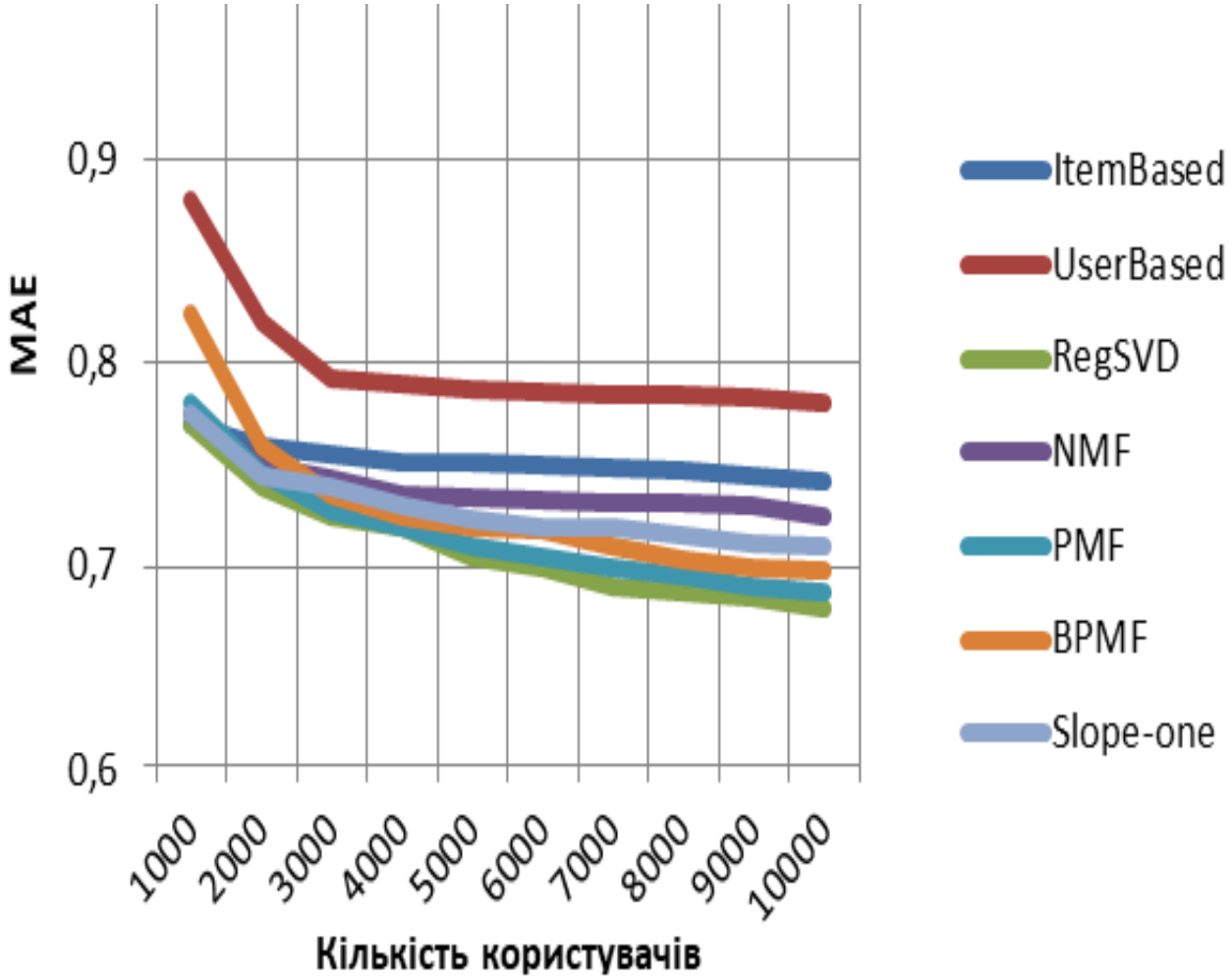

Рисунок 3.2 – графік залежності MAE (середня абсолютна похибка) від кількості користувачів у наборі данних

На рисунку 3.2 представлені графіки залежності середньої абсолютної похибки (MAE) від кількості користувачів.

З цього графіку ми можемо зробити наступні висновки:

1. Матричні методи факторизації, як правило, показують кращу продуктивність, коли кількість користувачів достатньо велика (> 3 000).

2. В цілому, найбільш ефективним алгоритмом є регуляризований SVD.

3. Коли кількість користувачів досить мала, між методами матричної факторизації та простими методами, орієнтованими на сусідстві, дуже мало відмінностей.

4. Регуляризований SVD, PMF, BPMF, як правило, найбільш чутливі до варіації кількості користувачів.

5. NMF нечутливий до кількості користувачів.

6. Існує суттєва різниця в чутливості між двома популярними методами item-base і user-base.

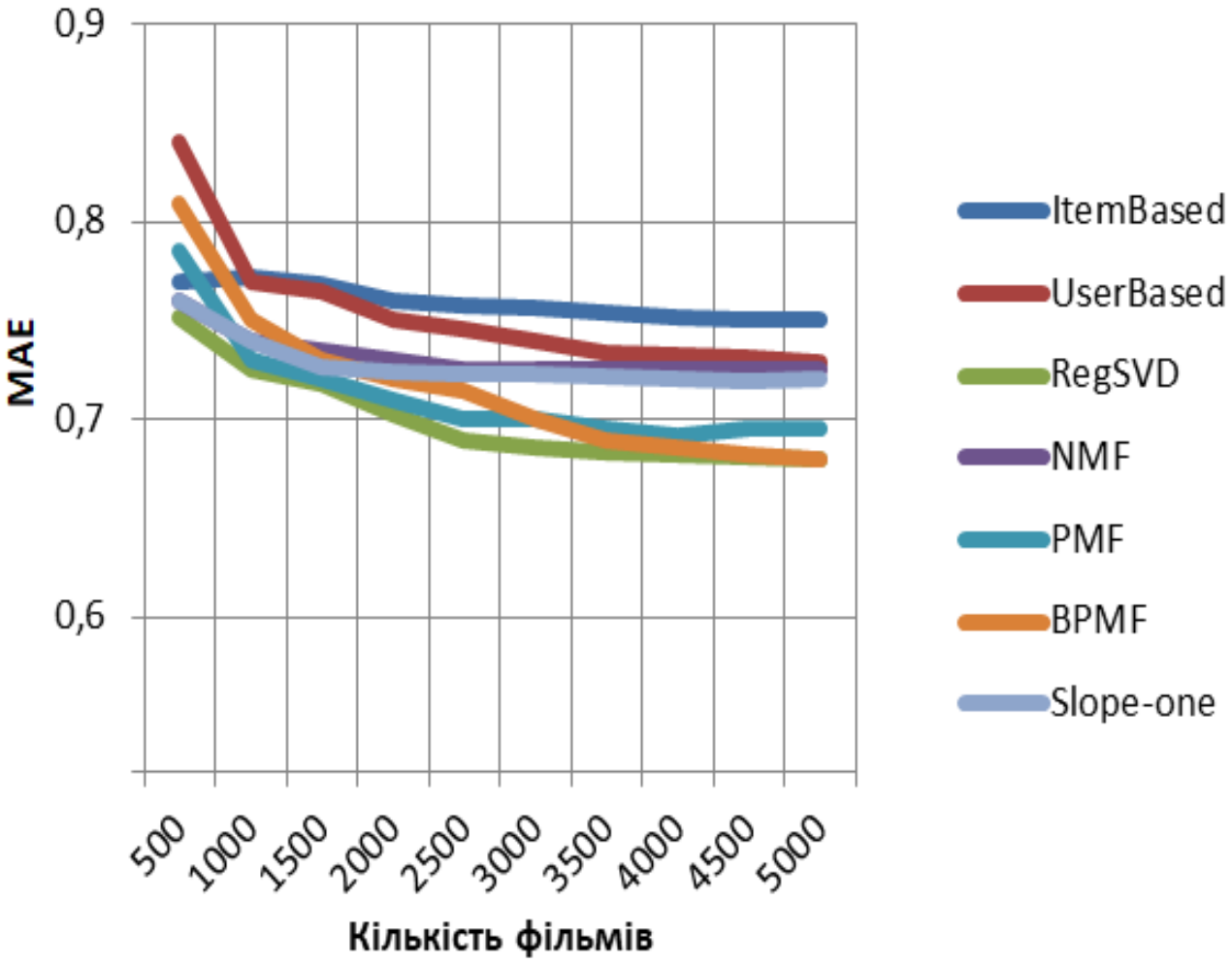

Рисунок 3.3 – графік залежності MAE (середня абсолютна похибка) від кількості фільмів у наборі данних

На рисунку 3.2 представлені графіки залежності середньої абсолютної похибки (MAE) від кількості.

З цього графіку ми можемо зробити наступні висновки:

1. Матричні методи факторизації, як правило, показують кращу продуктивність, коли кількість предметів(фільмів) стає достатньо великим (> 1 000).

2. В цілому, найбільш ефективним алгоритмом є регуляризований SVD.

3. Коли кількість користувачів досить мала, між методами матричної факторизації та простими методами, що базуються на сусідстві, існує дуже мало відмінностей.

4. User-based, регуляризований SVD, PMF, BPMF, як правило, є найбільш чутливим до зміни кількості предметів(фільмів).

5. Item-based і NMF, як правило, менш чутливі до зміни кількості предметів(фільмів).

6. Існує суттєва різниця в чутливості між двома популярними методами, орієнтованими на сусідство: user-based і item-based.

7. ItemBased КФ є надзвичайно ефективною при малій кількості елементів, але має майже постійну залежність від кількості елементів.

8. У поєднанні з спостереженнями до малюнку 3.2, можна зробити висновки, що схил-один і НМФ нечутливі до варіацій як для користувачів, так і для підрахунку позицій. slope-one і NMF відносно чутливі до варіацій як користувачів, так елементів.

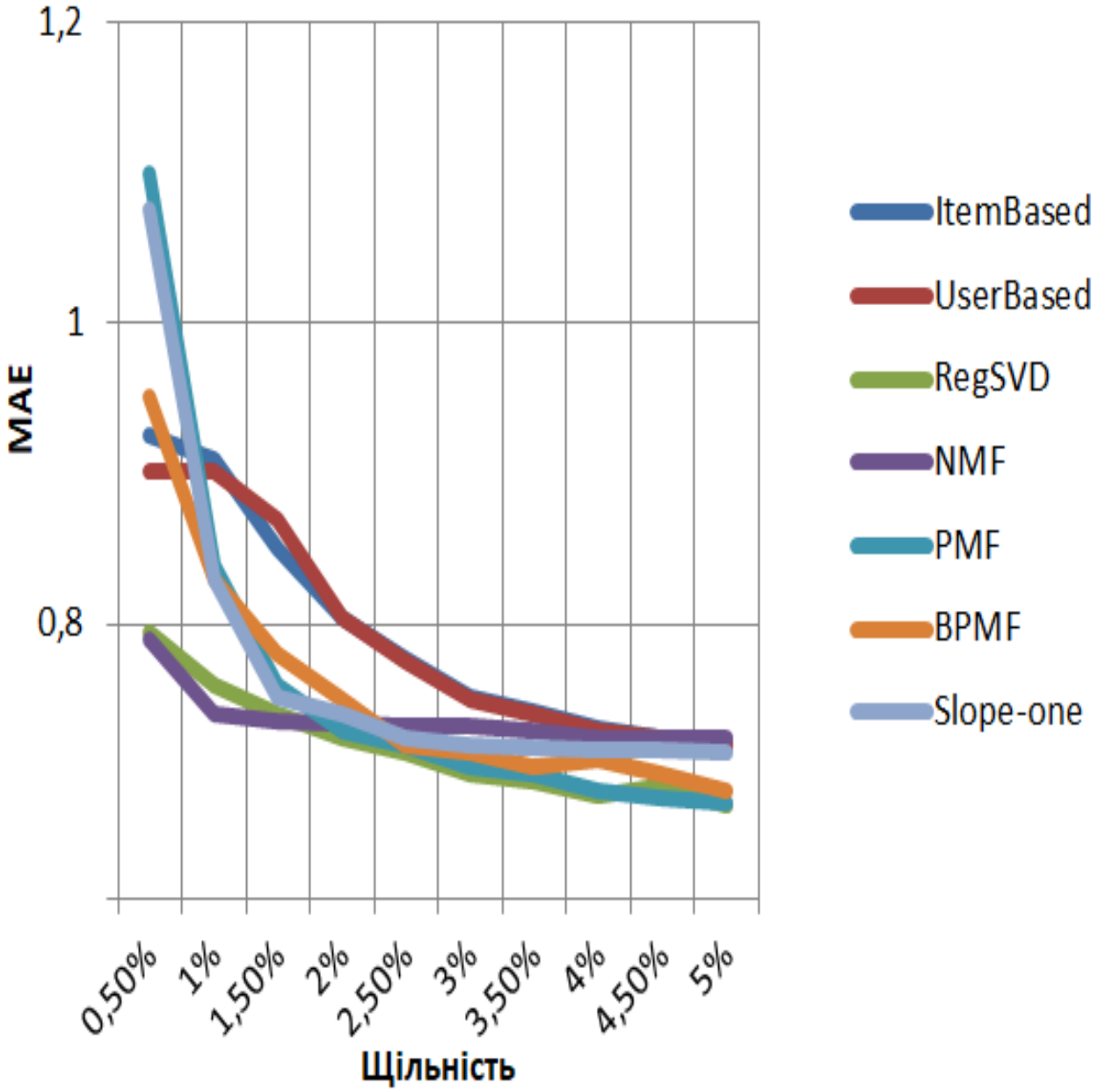

Рисунок 3.4 – графік залежності MAE (середня абсолютна похибка) від щільності набору даних

На рисунку 3.4 представлені графіки залежності середньої абсолютної похибки (MAE) від щільності дата сету.

З цього графіку ми можемо зробити наступні висновки:

1. Найкращим алгоритмом, як і у попередніх дослідах, виявився регуляризований SVD.

2. User-based, item-based, slope-one, PMF та BPMF показують сильну залежність від щільності.

6. NMF відносно незалежний від щільності.

7. Показники slope-one і PMF показують слабкі результати при більшій розрідженості, але при більш високій щільності вони дають досить пристойні результати.

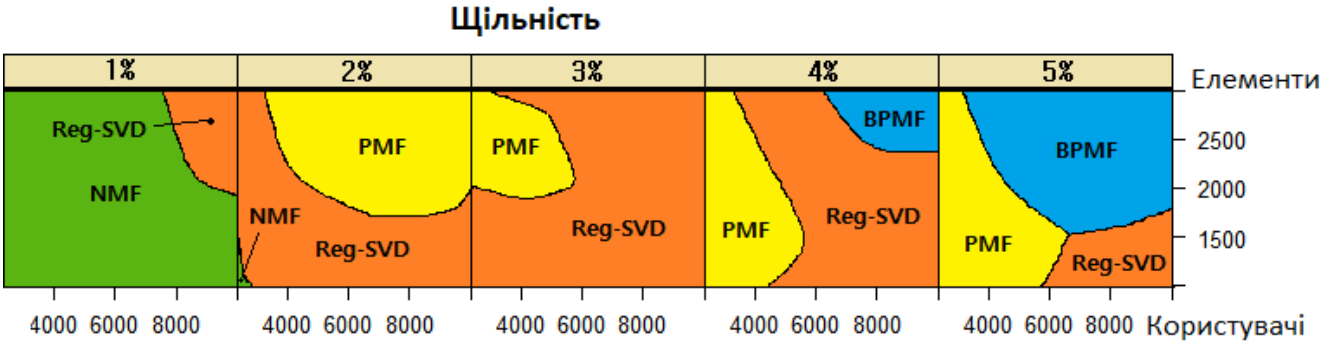

Рисунок 3.5 – найкращі алгоритми за MAE (середня абсолютна похибка) від кількості користувачів, елементів та щільності

З цього рисунку ми можемо зробити наступні висновки:

1. Ідентифікація найбільш ефективного алгоритму коливається, вона нелінійно залежить від кількості користувачів, кількості елементів та щільності.

2. NMF є домінуючим за низької щільності, тоді як BPMF добре працює для випадків з високою щільністю (особливо для багатьох користувачів та елементів).

3. Регуляризована SVD та PMF найкраще працюють для рівнів щільності 2% -4%.

3.2 Дослідження швидкодії алгоритмів

В цьому пункті ми покажемо залежність швидкодії алгоритмів від кількості користувачів, елементів та щільності. Час обрахунку включає в себе як час на обрахунок рекомендації так і час на навчання алгоритму (якщо воно необхідне).

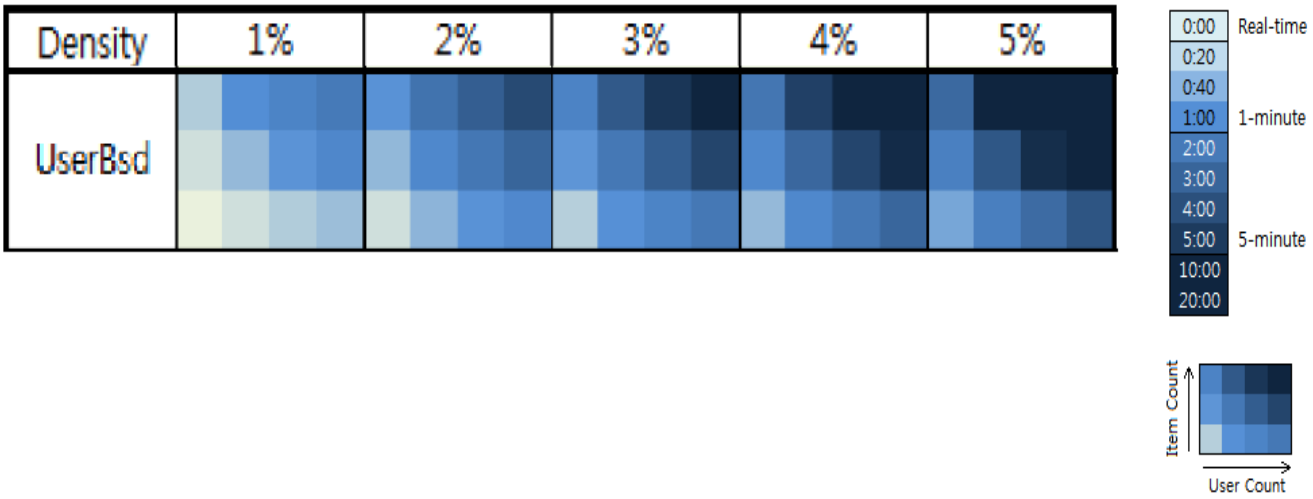

Рисунок 3.6 – залежність алгоритму User Based від кількості користувачів, елементів та щільності набору даних

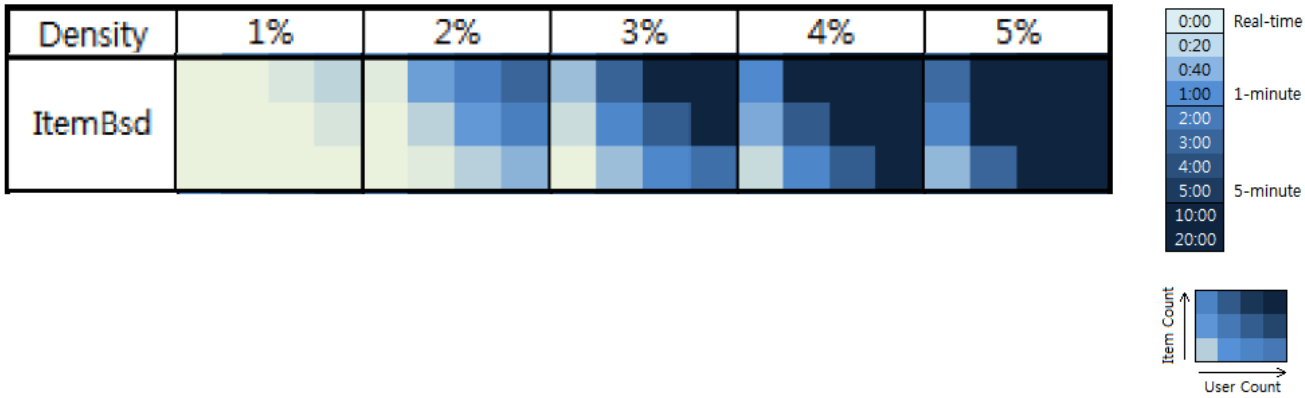

Рисунок 3.7 – залежність алгоритму Item Based від кількості користувачів, елементів та щільності набору даних

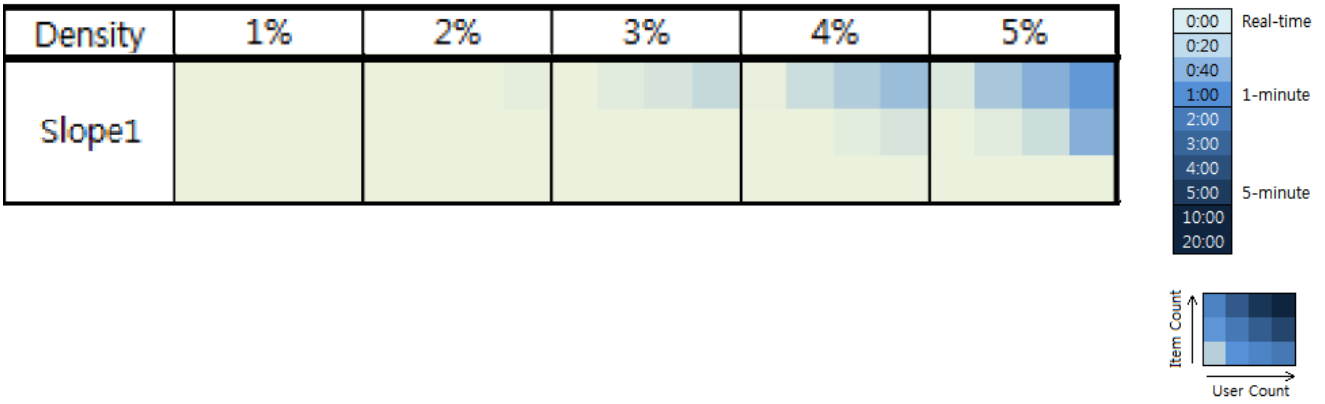

Рисунок 3.8 – залежність алгоритму Slope One від кількості користувачів, елементів та щільності набору даних

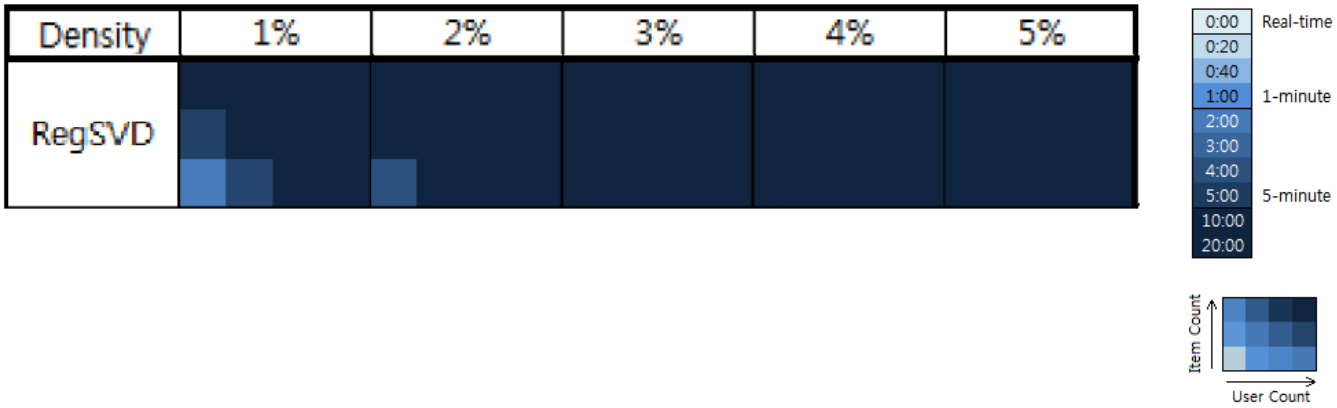

Рисунок 3.9 – залежність алгоритму Reg SVD від кількості користувачів, елементів та щільності набору даних

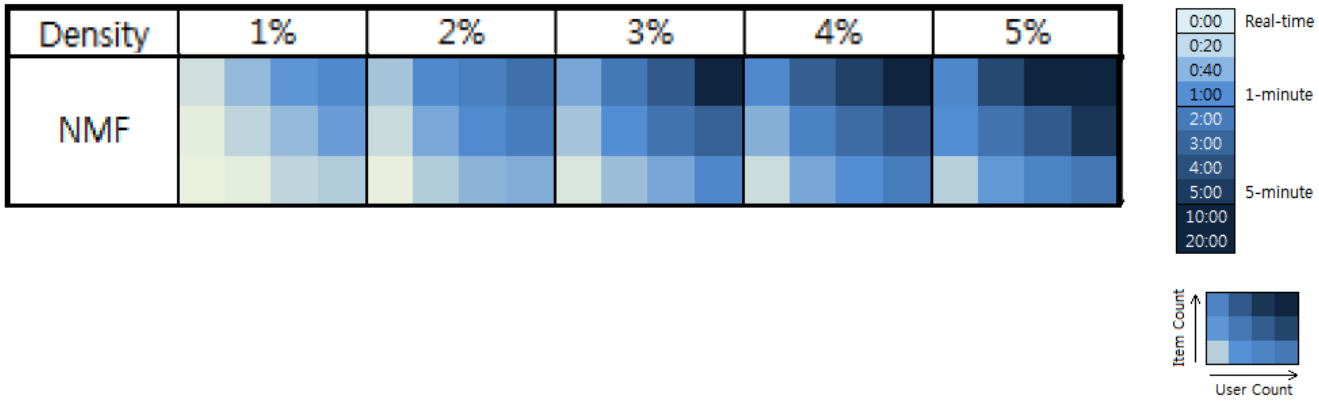

Рисунок 3.10 – залежність алгоритму NMF від кількості користувачів, елементів та щільності набору даних

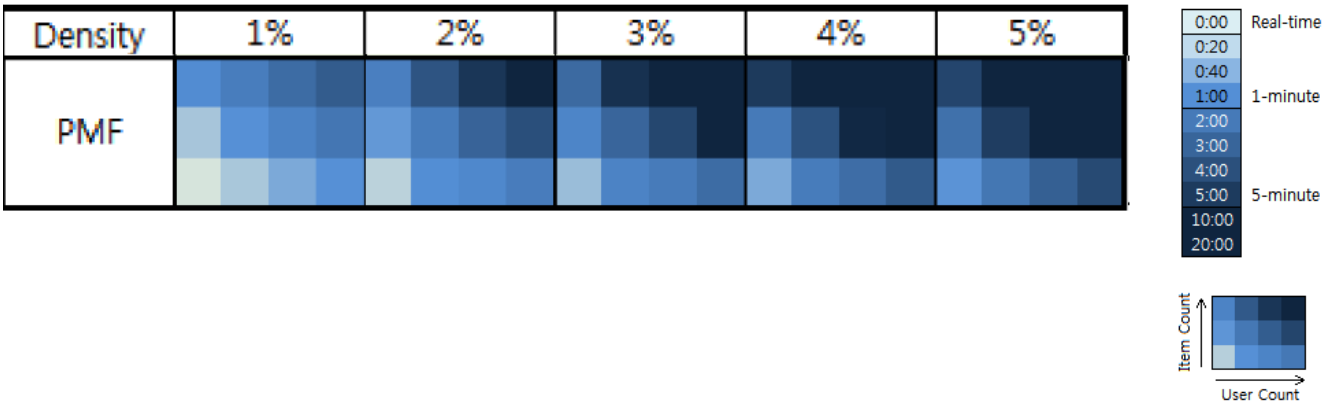

Рисунок 3.11 – залежність алгоритму PMF від кількості користувачів, елементів та щільності набору даних

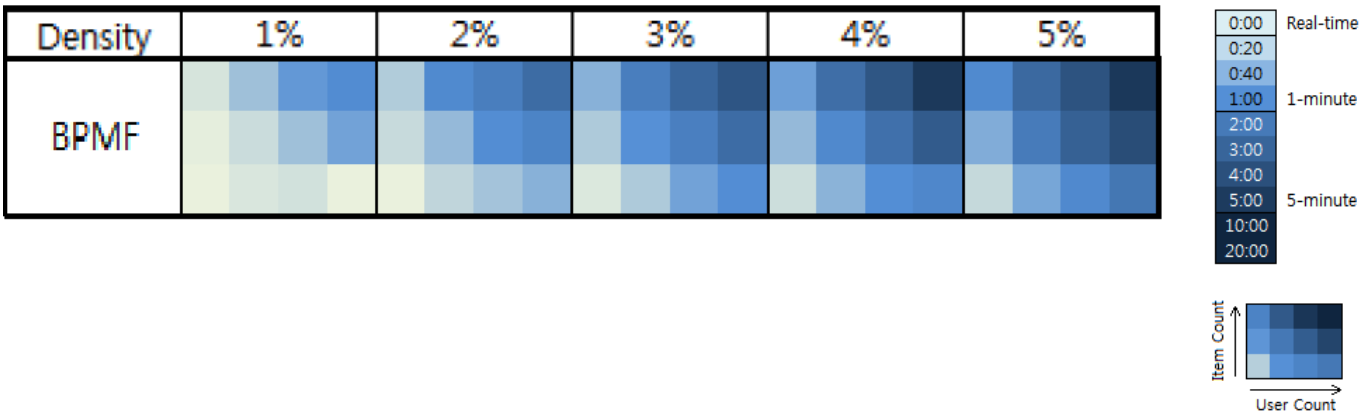

Рисунок 3.12 – залежність алгоритму BPMF від кількості користувачів, елементів та щільності набору даних

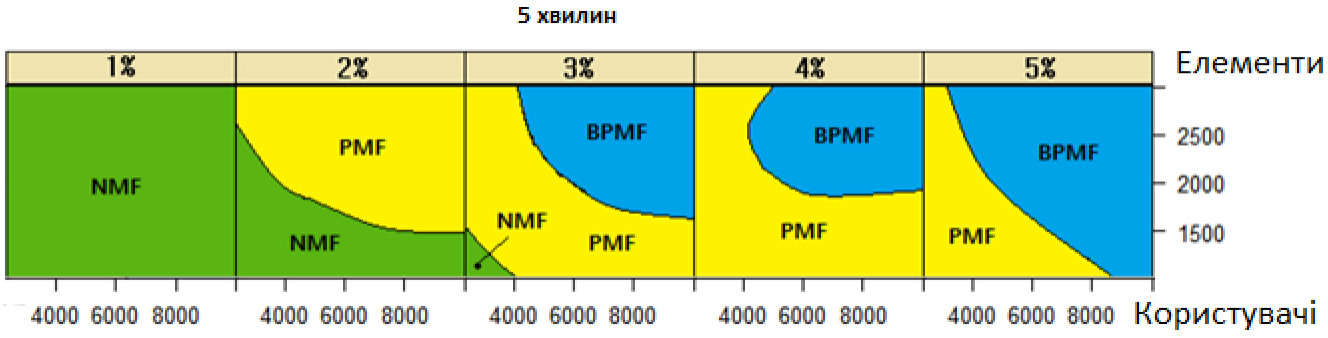

Рисунок 3.13 – залежність швидкодії алгоритму від кількості користувачів, елементів та щільності набору даних

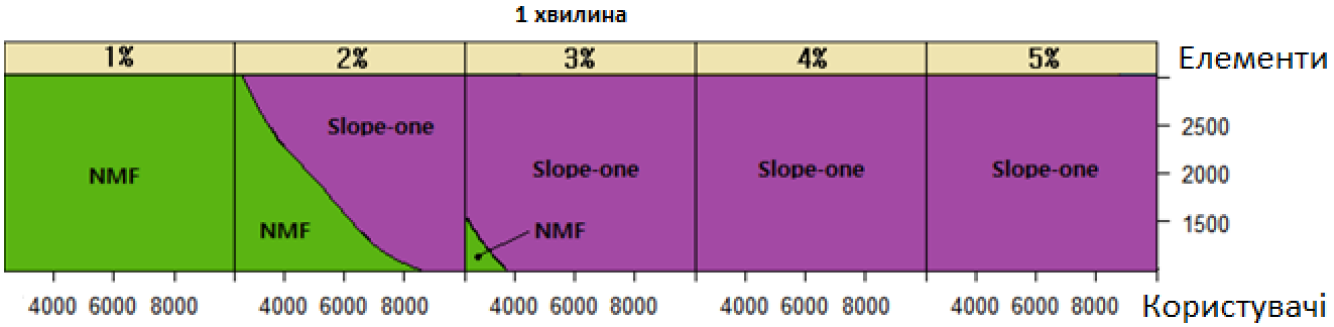

Рисунок 3.14 – залежність швидкодії алгоритму від кількості користувачів, елементів та щільності набору даних

Висновки по швидкодії алгоритмів:

1. Якщо в нас нема часових обмежень (маємо нескінченно часу на обчислення), застосовуються висновки з попереднього розділу. NMF найкраще працює для розріджених наборів даних, BPMF найкраще працює для щільних наборів даних, а PMF добре працює з меншим числом користувачів, тоді як Regularised SVD добре працює з меншою кількістю елементів.

2. Якщо обмеження часу становить 5 хвилин, не враховуються Регуляризоване SVD (на рисунку 3.8 темні кольори, тобто він потребує більше часу ніж ми можемо собі дозволити). У такому разі NMF найкраще працює для розріджених даних, BPMF найкраще працює для щільних і великих даних, а PMF найкраще працює в інших випадках.

3. Якщо обмеження часу становить 1 хвилину, PMF та BPMF також виключаються з розгляду через час, необхідний на їх роботу. Slope One найкраще працює в більшості випадків, за винятком найбільш розріджених даних, де NMF працює найкраще.

3.3 Рекомендації застосування алгоритмів КФ

Для початку давайте узагальнимо результати, що були отримані у попередньому пункті, у вигляді таблиці.

Таблиця 5.1 Характеристика алгоритмів

		User Based, Item Baser	Reg SVD, PMF, BPMF	NMF	Slope One
Залежність	Розмір	Низька	Висока	Низька	Середня
	Щільність	Висока	Дуже висока	Низька	Середня
Тостість	Щільні	Висока	Середня	Дуже висока	Задовільна
	Розріджені	Дуже погана	Дуже висока	Висока	Дуже висока
Обчислення	Тренування	Нема	Повільно	Швидко	Дуже швидко
	Обчислення	Дуже повільно	Середньо	Середньо	Швидко
Витрати пам’яті		Великі	Великі	Великі	Малі
Переваги		Не треба тренувати	Добре працює з щільним набором даних	Найкраще працює з розрідженими даними. Швидке навчання	Незважаючи на швидку роботу, дає гарний результат
Недоліки		Потребує багато пам’яті	Треба корегувати багато параметрівДовге обчислення	Треба корегувати багато параметрів.	Погано працює без великого / щільного набору даних

Отже, для застосування в реальному програмному забезпеченні треба враховувати наступні параметри:

- Скільки максимально часу ми можемо дати системі на обчислення
- Орієнтовну кількість користувачів та елементів
- Орієнтовну розрідженість даних

Виникає наступне питання. Як нам побудувати систему, щоб вона мала в собі сильні сторони перерахованих вище алгоритмів, та не мала їх недоліків.
Нашою пропозицією є комбінація алгоритмів. Ми, наприклад, можемо застосовувати в нашій системі одразу декілька алгоритмів. Нам буде необхідно аналізувати наш набір даних, визначати його розрідженість та об’єм. Для розріджених наборів даних можемо застосовувати NMF, для більш щільних – Reg SVG, але ці алгоритми потребують достатньо велику кількість часу на обчислення, тож ми можемо запускати їх (один з них, в залежності від розрідженості дата сету) декілька разів на добу, або з певним інтервалом. А для нових користувачів, які ще не мають рекомендацій, або «старих», які хочуть отримати більшу вибірку – застосуємо алгоритм Slope One, тому що він поєднує в собі дуже високу швидкість роботи та відмінну, як для такої швидкості) якість рекомендацій.

3.4 Висновки

В цьому розділі було наведено результати проведеного дослідження. Було наведено графіки з результатами дослідження точності алгоритмів в залежності від кількості користувачів, елементів та щільності набору даних.

Наведено дані стосовно часу необхідного алгоритму на надання рекомендації (час на навчання + час на надання рекомендації).

Показано, які алгоритми краще працюють при різній кількості користувачів, елементів та щільності набору даних. Наведено подібні результати для випадків, коли ми обмежені у часі і не можемо собі дозволити витрачати більше часу, ніж у нас є, на обчислення рекомендації.

По кожному з пунктів, наведених в цьому розділів, надано висновки стосовно швидкодії та якості надання рекомендацій алгоритмів.

4 РОЗРОБЛЕННЯ СТАРТАП-ПРОЕКТУ “MAGIC ASSISTANT”

Стартап як форма малого ризикового (венчурного) підприємництва впродовж останнього десятиліття набула широкого розповсюдження у світі через зниження бар’єрів входу в ринок (із появою Інтернету як інструменту комунікацій та збуту стало простіше знаходити споживачів та інвесторів, займатись пошуком ресурсів, перетинати кордони між ринками різних країн), і вважається однією із наріжних складових інноваційної економіки, оскільки за рахунок мобільності, гнучкості та великої кількості стартап-проектів загальна маса інноваційних ідей зростає.

Проте створення та ринкове впровадження стартап-проектів відзначається підвищеною мірою ризику, ринково успішними стає лише невелика частка, що за різними оцінками складає від 10% до 20%. Ідея стартап-проекту, взята окремо, не вартує майже нічого: головним завданням керівника проекту на початковому етапі його існування є перетворення ідеї проекту у працюючу бізнес-модель, що починається із формування концепції товару (послуги) для визначеної клієнтської групи за наявних ринкових умов. Розроблення та виведення стартап-проекту на ринок передбачає здійснення низки кроків, в межах яких визначають ринкові перспективи проекту, графік та принципи організації виробництва, фінансовий аналіз та аналіз ризиків і заходи з просування пропозиції для інвесторів.

4.1 Опис ідеї проекту

В межах підпункту було проаналізовано і подано у вигляді таблиць:

● зміст ідеї (що пропонується);

● можливі напрямки застосування;

● основні вигоди, що може отримати користувач товару (за кожним напрямком застосування);

● чим відрізняється від існуючих аналогів та замінників;

Таблиця 4.1 Опис ідеї стартап-проекту

Зміст ідеї	*Напрямки застосування*	*Вигоди для користувача*
Універсальна система рекомендацій, яка включає в себе рекомендації новин, фільмів, музики та книг базуючись на особистих вподобаннях, попередніх виборах, пошукових запитах та аналізі тенденцій інших користувачів.	1. Для пошуку новин 2. Для пошуку потенційно цікавих фільмів 3. Для пошуку потенційно цікавих пісень 4. Для потенційно цікавих книг	1. Багато контенту в одному місці 2. Висока якість рекомендацій 3. Зручний інтерфейс 4. Гнучке налаштування

У таблиці 4.1 було описано ідею стартапу «Magic Assistant». Це має бути універсальний веб сервіс з надання рекомендацій в багатох галузях, а саме: рекомендації новин, фільмів, музики та книг. Цей веб сервіс має великий потенціал в зацікавленні користувачів через об’єднання великої кількості можливостей в одному місці, потужному алгоритмі рекомендацій та простоті у користуванні та переході поміж розділами сервісу.

Для реалізації проекту необхідно визначити його характеристики, сильні та слабкі сторони в порівнянні з конкурентами.

Таблиця 4.2 Визначення сильних, слабких та нейтральних характеристик ідеї проекту

№	*Техніко-економічні характеристики ідеї*	*(Потенційні) товари/концепції конкурентів*				*W (слабка сторона)*	*N (нейтральна сторона)*	*S (сильна сторона)*
		Мій проект	*Конкурент 1*	*Конкурент 2*	*Конкурент3*			
1	Кросплатформеність	Так	Так	Так	Ні			+
2	Форма виконання	Веб-сервіс + Мобільний додаток	Веб-сервіс	Веб-сервіс + Мобільний додаток	Сервіс			+
3	Собівартість	Середня	Низька	Висока	Середня		+	
4	Зручність використання	Висока	Низька	Середня	Висока			+
5	Складність реалізації	Висока	Середня	Висока	Висока		+	

До сильних сторін даного проекту можна віднести його кросплатформенність, що дозволить більшій кількості користувачів скористатись

ним на зручному для них девайсі та улюбленій операційній системі. Також він має більш зручний інтерфейс з можливістю налаштування на смак користувача. Реалізація проекту в двох варіантах (веб-додаток та мобільний додаток) дозволить користувачам мобільних девайсів користуватись нашим додатком з більшою простотою та швидкістю завдяки зменшеним вимогам до інтернет трафіку. Інші характеристики є нейтральними, тож даний проект можна вважати конкурентоспроможним.

4.2 Технологічний аудит ідеї проекту.

Визначення технологічної здійсненності ідеї проекту передбачає аналіз таких складових:

- за якою технологією буде виготовлено товар згідно ідеї проекту?
- чи існують такі технології, чи їх потрібно розробити/доробити?
- чи доступні такі технології авторам проекту?

Для реалізацій ідеї необхідно перевірити, чи можна втілити цей проект у життя.

Таблиця 4.3 Технологічна здійсненність ідеї проекту

№	*Ідея проекту*	*Технології її реалізації*	*Наявність технологій*	*Доступність технологій*
1	Magic Assistant	Java	Є у наявності	Доступно на більшості платформах. Безкоштовна.
		Spring	Є у наявності	Доступно на більшості платформах. Безкоштовна.
		React Native	Є у наявності	Доступно на більшості платформах. Безкоштовна.
		React	Є у наявності	Доступно на більшості платформах. Безкоштовна.
		Redux	Є у наявності	Доступно на більшості платформах. Безкоштовна.
		Webpack	Є у наявності	Доступно на більшості платформах. Безкоштовна.
Обрана технологія реалізації ідеї проекту: Java + Spring + React Native + React + Redux + Webpack				

У таблиці 4.3 було наведено перелік технологій, які будуть використані для реалізації даного проекту. Для написання серверної частини та REST API буде використана Java та її фреймворк Spring. Для написання клієнтсьогоко інтерфейсу веб-сервісу буде використаний React + Redux. Мобільний додаток для iOS та Android буде реалізовано за допомогою React Native + Redux.

За результатами аналізу таблиці зроблено висновок, про можливість реалізації проекту.

4.3Аналіз ринкових можливостей запуску стартап-проекту.

Визначення ринкових можливостей, які можна використати під час ринкового впровадження проекту, та ринкових загроз, які можуть перешкодити реалізації проекту, дозволяє спланувати напрями розвитку проекту із урахуванням стану ринкового середовища, потреб потенційних клієнтів та пропозицій проектів-конкурентів.

Спочатку було проведено аналіз попиту: наявність попиту, обсяг, динаміка розвитку ринку (табл. 4.4).

Таблиця 4.4 Попередня характеристика потенційного ринку стартап-проекту

№	*Показники стану ринку (найменування)*	*Характеристика*
1	Кількість головних гравців, од	5
2	Загальний обсяг продаж, грн/ум.од	1000 грн./ум.од
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу (вказати характер обмежень)	Немає
5	Специфічні вимоги до стандартизації та сертифікації	Немає
6	Середня норма рентабельності в галузі , %	R = 25%

Середню норму рентабельності в галузі було порівняно із банківським відсотком на вкладення. Останній є меншим, тому є сенс вкладати гроші саме у цей проект.

За результатами аналізу таблиці було зроблено висновок, що ринок є привабливим для входження.

Надалі були визначені потенційні групи клієнтів, їх характеристики, та зформовано орієнтовний перелік вимог до товару для кожної групи.

Таблиця 4.5 Характеристика потенційних клієнтів стартап-проекту

№	*Потреба, що формує ринок*	*Цільова аудиторія (цільові сегменти ринку)*	*Відмінності у поведінці різних потенційних цільових груп клієнтів*	*Вимоги споживачів до товару*
	Потреба в полегшенні пошуку контенту, та підвищення якості рекомендуємого контенту.	Активні користувачі інтернету різних вікових категорій	Цільова група, що хоче поширити свої товари за допомогою реклами	Рішення має бути крос-платформеним та інтуїтивно-зрозумілим для використання. Має надавати якісні рекомендації.

В таблиці 4.5 було описано цільову аудиторію даного проекту. Наша цільова аудиторія дуже велика, адже інтернетом користується величезна кількість людей і майже всі цікавляться хоча б однією з можливостей даного проекту. Даний проект може зацікавити їх якістю рекомендацій та зручним інтерфейсом.

Після визначення потенційних груп клієнтів було проведено аналіз ринкового середовища: складено таблиці факторів, що сприяють ринковому впровадженню проекту, та факторів, що йому перешкоджають (табл. 4.6-4.7).

Таблиця 4.6 Фактори загроз

№	*Фактор*	*Зміст загрози*	*Можлива реакція компанії*
	Конкуренція	Вихід на ринок одного з гігантів сумісних областей з комплексним програмним рішенням, що міститиме у собі аналог нашого продукту	1. Передбачити додаткові переваги власного проекту для того, щоб повідомити про них саме після виходу міжнародної компанії на ринок. 2. Обрати нову цільову аудиторію і зосередитися на ній 3. Об'єднання з компанією-конкурентом
	Економічний	Подорожчення вартості та обслуговування обладна, необхідного для роботи системи	1. Оптимізація програмного продукту, для можливості його запуску на більш бюджетних пристроях.
	Зміна потреб користувачів	Користувачам необхідний інший функціонал	1. Передбачити можливість дозширення функціоналу
	Законодавчі	Зміни в законодавстві стосовно обробки персональних послуг користувачів.	1. Впровадження контроль за збереженням персональних даних користувачів
	Віруси	Зараження серверів комп'ютерними вірусами	1. Використання противірусних програм 2. Використання більш безпечних операційних систем 3. Створення резервних копій даних

В таблиці вище були наведені фактори загроз та способи зменшення ризиків. Найбільшою загрозою є конкуренція. Для боротьби з конкуренцією нам необхідно передбачити найкращий набір функціоналу. Також необхідно передбачити можливість додавання нового функціоналу, або зробити фокус на більш вузькій цільовій аудиторій, надаючи їм більше уваги. Також, при виході на риноу дуже

потужного конкурента, можна розглянути можливість об’єднання або поглинання з метою збереження вкладених в даний проект коштів.

Таблиця 4.7 Фактори можливостей

№	*Фактор*	*Зміст можливості*	*Можлива реакція компанії*
	Науково-технічний	Тенденція до випуску покращеного спеціалізованого обладнання та розробка більше ефективних алгоритмів	Адаптація існуючого рішення і алгоритмів під нову технологію.
	Попит	Більш широке розповсюдження технології рекомендаційних систем	Постійна підтримка продукту.
	Зростання можливостей потенційних покупців	Зростання фінансування у підприємств, яким може знадобитись реклама, або підвищення рівня життя у користувачів	1. Запропонувати підприємствам розмістити рекламу на нашому сервісі. 2. Розроблення додаткових послуг для VIP користувачів
	Зниження довіри до конкурента 2	У додатку конкурента 2 нещодавно стався збій і протягом декількох днів додаток працював зі збоями	Звертати увагу клієнтів на надійність нашого сервісу
	Запуск мобільного 4G	Збільшення швиткості мобільного інтернету	Звернутись до провайдерів мобільного інтернету щодо надання користувачам пільгових тарифів на користування нашим сервісом.

В цій таблиці було розглянуто фактори можливостей. Найбільш цікавим для нас звичайно є зниження довіри до конкурентів, адже це автоматично піднімає наші позиції серед конкурентів. Також вагомим є фактор збільшення можливостей покупців, адже це збільшить кількість потенційних клієнтів, які захочуть замовити у нас рекламу, що в свою чергу підніме її вартість.

Надалі було проведено аналіз пропозиції: визначили загальні риси конкуренції на ринку (табл. 4.8).

Таблиця 4.8 Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	*В чому проявляється дана характеристика*	*Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)*
1. Вказати тип конкуренції: Нецінова конкуренція.	Існує декілька конкурентів, які повторюють деякі функції нашої системи	Підтримка якості продукту та постійні нововведення
2. За рівнем конкурентної боротьби: Міжнародний.	Фірми-конкуренти знаходяться як в нашій країні так і в інших країнах.	Адаптація продукту як для вітчизняних так і для зарубіжних клієнтів.
3. За галузевою ознакою: внутрішньогалузева.	Продукт використовується лише всередині даної галузі.	Постійне вдосконалення продукту.
4. Конкуренція за видами товарів: товарно-родова.	Системи конкурентів виконують подібні функції але досить відрізняються він нашої	Створити продукт, врахувавши сильні і слабкі сторони конкурентів.
5. За характером конкурентних переваг: нецінова.	Збільшення функціональності в межах однієї системи та збільшення якості її роботи	Зниження ціни на продукт та підтримка його якості.
6. За інтенсивністю: марочна.	Бренди існують і конкурують.	PR, реклама, просування бренду.

В цій таблиці наведено аналіз конкуренції на ринку. Визначено, що най сервіс працює в середовищі нецінової, товарно-родової конкуренції, адже наші конкуренти частково повторюють функціонал нашого сервісу, тож конкуренція ведеться за рахунок якості надання послуг. Конкуренція відбувається не лише в середині країни, а й на міжнародному ринку надання послуг.

В наступній таблиці буде проведено аналіз конкуренції в галузі за М. Портером.

Таблиця 4.9 – Аналіз конкуренції в галузі за М. Портером

Складові аналізу	*Прямі конкуренти в галузі*	*Потенційні конкуренти*	*Постачальники*	*Клієнти*	*Товари-замінники*
	Навести перелік прямих конкурентів	*Визначити бар'єри входження в ринок*	*Визначити фактори сили постачальників*	*Визначити фактори сили споживачів*	*Фактори загроз з боку замінників*
Висновки	Існує 3 конкуренти на ринку. Найбільш схожим за виконанням є конкурент 2, так як його рішення також представлене у двох варіантаї: Веб сервісу та мобільного додатку.	Так, можливості для входу на ринок є, бо наше рішення поєднує в собі велику кількість можливостей, а також має зручний інтерфейс та являється кросплатформенним	Постачальники відсутні	Важливим для користувача є зручність у користуванні	Товари-замінники можуть використати більш дешеву технологію створення ПЗ та зменшити собівартість товару

В цій таблиці було досліджено аналіз конкуренції в галузі за М.Портером. Було визначено конкурента, чий продукт найбільш подібний до нашого. Також було визначено та обґрунтовано можливості виходу нашого сервісу на ринок, наведено його основні переваги та фактори сили споживачів. Також було розглянуто фактори загрози з боку конкурентів.

В наступній таблиці будуть наведені обґрунтування факторів конкурентоспроможності для нашого сервісу.

Таблиця 4.10 - Обґрунтування факторів конкурентоспроможності

№	*Фактор конкурентноспроможності*	*Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)*
1.	Використання ПЗ у вигляді веб-сервісу та мобільного додатку	Зручне використання сервісу можливе з великої кількості пристроїв. Кожен з користувачів може відкрити даний сервіс у зручний для себе спосіб з максимальною зручністю.
2.	Простота інтерфейсу користувача	Інтуїтивно зрозумілий інтерфейс з простим доступом до найважливіших функцій даного сервісу.

В таблиці вище наведені фактори конкурентоспроможності, а також обґрунтування до них. Головними факторами конкурентоспроможності даного сервісу є те, що він існує як в варіанті веб-сервісу так і у варіанту мобільного додатку. Також перевагою нашого сервісу над конкурентами є дуже зручний та простий інтерфейс, за допомогою якого можна за мінімальний час отримати потрібний результат.

В таблиці нижче наведено порівняльний аналіз сильних та слабких сторін проекту.

Таблиця 4.11 - Порівняльний аналіз сильних та слабких сторін проекту

№	*Фактор конкурентноспроможності*	*Бали 1-20*	*Рейтинг товарів-конкурентів у порівнянні з нашим підприємством*						
			-3	*-2*	*-1*	*0*	*+1*	*+2*	*+3*
1.	Реалізація у вигляді веб-сервісу та мобільного додатку	18			+				
2.	Простота інтерфейсу користувача	20	+						

В цій таблиці наведено порівняльний аналіз сильних та слабких сторін даного проекту в порівнянні з проектами конкурентів. Можна зробити висновок, що даний проект є досить перспективним, тому що він не поступається конкурентам, а іноді й перевершує їх, як наприклад у простоті інтерфейсу.

В таблиці 4.12 наведено SWOT-аналіз стартап-проекту, базуючись на характеристиках проекту, що були надані в попередніх таблицях.

Таблиця 4.12 – SWOT-аналіз стартап-проекту

Сильні сторони: Зручне використання сервісу можливе з великої кількості пристроїв. Кожен з користувачів може відкрити даний сервіс у зручний для себе спосіб з максимальною зручністю. Інтуїтивно зрозумілий інтерфейс з простим доступом до найважливіших функцій даного сервісу.	Слабкі сторони: висока складність реалізації проекту через необхідність підключення великої кількості сторонніх сервісів.
Можливості: Зростання фінансування у підприємств, яким може знадобитись реклама, або підвищення рівня життя у користувачів. У додатку конкурента 2 нещодавно стався збій і протягом декількох днів додаток працював зі збоями. Збільшення швиткості мобільного інтернету. Збільшення швиткості мобільного інтернету. Тенденція до випуску покращеного спеціалізованого обладнання та розробка більше ефективних алгоритмів.	Загрози: Вихід на ринок одного з гігантів сумісних областей з комплексним програмним рішенням, що міститиме у собі аналог нашого продукту. Подорожчення вартості та обслуговування обладна, необхідного для роботи системи. Користувачам необхідний інший функціонал. Зміни в законодавстві стосовно обробки персональних послуг користувачів. Зараження серверів комп'ютерними вірусами.

В таблиці 4.12 було наведено SWOT-аналіз нашого стартап-проекту. Описано сильні сторони, такі як зручний інтерфейс та особливості реалізації, описано головний недолік, а саме високу складність реалізації. Надано інформацію про можливості та загрози. Основними можливостями є невдачі конкурентів та збільшення фінансових можливостей у потенційних замовників реклами. Основними загрозами традиційно є конкуренція та зміни потреб користувачів.

В таблиці 4.13 наведено альтернативи ринкового впровадження сатартап-проекту.

№	*Альтернатива (орієнтовний комплекс заходів) ринкової поведінки*	*Ймовірність отримання ресурсів*	*Строки реалізації*
1.	Створення ПЗ використовуючи нейронні мережі	80%	18 місяців
2.	Створення ПЗ на основі класичних методів машинного навчання	30%	24 місяців

В таблиці вище наведено альтернативні варіанти реалізації стартап-проекту. Наведено відсоток ймовірності отримання позитивного результату та приблизні терміни реалізації. З зазначених альтернатив перевага надається тій, яка має більшу ймовірність отриманна результату та має більш стислий термін реалізаці. Тож ми обираємо першу альтернативу.

4.4 Розробка ринкової стратегії проекту

Розроблення ринкової стратегії першим кроком передбачає визначення стратегії охоплення ринку: опис цільових груп потенційних споживачів.

Таблиця 4.14 - Вибір цільових груп потенційних споживачів

№	*Опис профілю цільової групи потенційних клієнтів*	*Готовність споживачів сприйняти продукт*	*Орієнтовний попит в межах цільової групи (сегменту)*	*Інтенсивність конкуренції в сегменті*	*Простота входу у сегмент*
1.	Підприємства	Можливість розміщення реклами	Середній	Існує 3 конкуренти, які надають схожі, але менш швидкі та менш результативні рішення.	Швидкодія, зручний користувацький інтерфейс
2.	Інтрернет-користувачі	Пошук цікавого контенту	Великий		Швидкодія, зручний користувацький інтерфейс
Які цільові групи обрано: обираємо інтернет-користувачі та підприємства					

В таблиці 4.14 було описано основні цільові групи потенційних споживачів. Було наведено орієнтовний попит в межах цільової групи, описано інтерсивність конкуренції в сегменту та орієнтовну складність виходу даного проекту на ринок. Основними цільовими групами було обрано підприємсстрав, які можуть бути зацікавлені в рекламі у нашому сервісі а також інтернет-користувачі, які можуть бути зацікавлені в одному джерелі великої кількості інформацій та розважального контенту.

В наступній таблиці буде описано базові стратегії розвитку

Таблиця 4.15 – Визначення базової стратегії розвитку

№	*Обрана альтернатива розвитку проекту*	*Стратегія охоплення ринку*	*Ключові конкурентоспромо жні позиції відповідно до обраної альтернативи*	*Базова стратегія розвитку*
1.	Створення ПЗ використовуючи нейронні мережі	Ринкове позиціонування	Швидкодія, простота у користуванні, реалізація у двох варіантах (веб-сервіс та мобільний додаток)	Диференціація

В вищезазначеній таблиці зазначні базові стратегії розвитку. Була отрана альтернатива розвитку проекту з використанням нейронних мереж. Стратегією охоплення ринку є ринкове позиціонування. Базовою стратегією розвитку є диференціація. Зазначеноосновні конкурентоспроможні позиції, а саме зручність простого інтерфейсу та наявність сервісу у двох варіантах.

В таблиці 4.16 буде визначено базову стратегію конкурентної поведінки.

Таблиця 4.16 - Визначення базової стратегії конкурентної поведінки

№	*Чи є проект «першопрохідцем» на ринку?*	*Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?*	*Чи буде компанія копіювати основні характеристики товару конкурента, і які?*	*Стратегія конкурентної поведінки*
1.	Ні	Так	Буде, а саме: основною задачею є розробка ПЗ з використанням нейронних мереж (конкуренти 1, 2, 3), зручний інтерфейс користувача (конкурент 3), Реалізація у вигляді веб сервісу та мобільного додатку (конкурент 2)	Зайняття конкурентної ніші

В таблиці 4.16 було визначено базові стратегії конкурентної поведінки. Наш проект не є «першопрохідцем» , але є всі передумови для заняття свої ніші на ринку. Даний проект буде намагатись перетягнути частину користувачів від конкурентів. Також ми плануємо взяти найбільш вдалі ріщення від проектів конкурентів з подальшим їх вдосконаленням.

В таблиці 4.17 буде описано визначення стратегії позиціонування

Таблиця 4.17 - Визначення стратегії позиціонування

№	*Вимоги до товару цільової аудиторії*	*Базова стратегі я розвитку*	*Ключові конкурентоспромо жні позиції власного стартап-проекту*	*Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)*
1.	Простота інтерфейсу, швидкодія, точність результатів	Диферен-ціація	Простота користувацького інтерфейсу дозволить отримувати необхідні дані і відслідковувати події в режимі реального часу	Швидкодія, простота, якість рекомендацій

В таблиці вище описано старатегію позиціонування даного стартап-проекту. Описано основні вимоги цільової аудиторії до товару, а саме простоту у користуванні та точність рекомендацій. Визначено базову стратегію розвитку (диференціація). Сформовано перелік ключових позицій конкурентоспроможності даного проекту.

4.5 Розробка маркетингової програми

Першим кроком є формування маркетингової концепції товару, який отримає споживач. Для цього у табл. 4.18 потрібно підсумувати результати попереднього аналізу конкурентоспроможності товару.

Таблиця 4.18 - Визначення ключових переваг концепції потенційного товару

№	*Потреба*	*Вигода, яку пропонує товар*	*Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити)*
1.	Реаліцяція як у вигляді веб сервісу так і у вигляді мобільного додатку	ПЗ працює однаково швидко та має однаково зручний інтерфейс як на ком'єютеках так і на мобільних пристроях	Перевага у зручності
2.	Кросплатформенність	ПЗ однаково успішно працює на різних платформах та операційних системах	Перевага у зручності

В таблиці 4.18 наведені ключві переваги концепції нашого проекту, а саме те, що проект являється кросплатформенним та реалізований у двох варіантах, таких як веб-сервіс та мобільний додаток (вони використовують одне API). Основною вигодою для користувача є те, що додаток однаково добре працює на різних пристроях та системах, що забезпечує його перевагу у зручності використання перед його конкурентами.

У наступній таблиці показана тритівнема маркетингова модель нашого проекту.

Таблиця 4.19 - Опис трьох рівнів моделі товару

<table>
<tr><td>Рівні товару</td><td colspan="3">Сутність та складові</td></tr>
<tr><td>I. Товар за задумом</td><td colspan="3">Універсальна система рекомендацій, яка включає в себе рекомендації новин, фільмів, музики та книг</td></tr>
<tr><td rowspan="5">II. Товар у реальному виконанні</td><td>Властивості/характеристики</td><td>М/Нм</td><td>Вр/Тх /Тл/Е/Ор</td></tr>
<tr><td>1. Зручність та простота користувацького інтерфейсу
2. Якість рекомендацій
3. Кросплатформенність</td><td>Не матеріальна</td><td>Технологічна</td></tr>
<tr><td colspan="3">Якість: згідно до стандарту ISO 4444 буде проведено тестування</td></tr>
<tr><td colspan="3">Маркування відсутнє</td></tr>
<tr><td colspan="3">Моя компанія: “ThunderSoft”</td></tr>
<tr><td rowspan="2">III. Товар із підкріплюнням</td><td colspan="3">3-місячна пробна VIP версія</td></tr>
<tr><td colspan="3">Постійна підтримка для користувачів</td></tr>
<tr><td colspan="4">За рахунок чого потенційний товар буде захищено від копіювання: наш сервіс буде захищений ліцензією.</td></tr>
</table>

В таблиці 4.19 було описано три рівні моделі товару. Було описано основний задум товару, а саме, що це має бути універсальна система рекомендацій. Було перераховано основні властивості продукту та його характеристики. Зазначено, що новим користувачам на 3-місячний період надається з додатковими послугами.

В таблиці 4.20 показано визначення цінових меж, якими необхідно керуватись при встаноовленні ціни на товар.

Таблиця 4.20 Визначення меж встановлення ціни

№	*Рівень цін на товари-замінники*	*Рівень цін на товари-аналоги*	*Рівень доходів цільової групи споживачів*	*Верхня та нижня межі встановлення ціни на товар/послугу*
	500-1000$	700-1200$	10000$	400-800$

Наступним кроком є визначення оптимальної системи збуту, в межах якого було прийняте рішення (табл. 4.20):

- проводити збут власними силами і залучати сторонніх посередників.
- користуватися однорівневим каналом збуту;

Таблиця 4.20 Формування системи збуту

№	*Специфіка закупівельної поведінки цільових клієнтів*	*Функції збуту, які має виконувати постачальник товару*	*Глибина каналу збуту*	*Оптимальна система збуту*
	Підписка на додаткові послуги. Замовлення реклами	Продаж	Однорівневий	Власні сили та через посередників

Останнім пунктом є таблиця 4.22, яка описує концепцію маркетингових комунікацій.

Таблиця 4.21 Концепція маркетингових комунікацій

№	*Специфіка поведінки цільових клієнтів*	*Канали комунікацій, якими користуються цільові клієнти*	*Ключові позиції, обрані для позиціонування*	*Завдання рекламного повідомлення*	*Концепція рекламного звернення*
	Клієнти обиратимуть цікаві їм новини, фільми, книги.	Соціальні мережі, електронна пошта, мобільні телефони	Ціна, простота використання, кросплатформеність, більш зручне та гнучке налаштування , висока якість рекомендацій	Показати переваги продукту, ефективність роботи для великої кількості випадків.	Демо ролик з використанням, реклама.

У таблиці вище описано концепцію маркетингових комунікацій даного проекту. Описано специфіку поведінки цільових клієнтів, канали комунікації для отримання зворотного зв’язку. Описано ключові пункти, що характеризують наш проект. Наведено завдання рекламного повідомлення та концепцію рекламного звернення.

5.7 Висновки

В даному розділі було проведено аналіз програмного продукту у якості стартап проекту. Можна зазначити що у проекта є можливість комерціалізації, адже ринок надання послуг в мережі інтернет з використанням рекомендаційних систем динамічно розвивається, створюються нові додатки які, в свою чергу, стимулють попит на різноманітні допоміжні засоби для пришвидшення роботи та оптимізації алгоритмів та матеріального забезпечення.

Було проведено аналіз ризиків та можливостей які можуть виникнути. Основними загрозами, очікувано, виявились конкуренція та зміна потреб користувачів. Найбільш вдалими можливостями для нас, звичайно ж, є невдачі наших конкурентів. Також гарною можливістю для росту є загальне «підняття» ринку.

На ринку наявна нецінова конкуренція, існує декілька фірм-конкурентів, але всі вони покривають лише якусь певну частину функціональності нашої системи, тому вихід на нього буде потребувати певних зусиль та капіталовкладень. Проте проект є доволі конкурентноспроможним завдяки своїй нижчій собівартості та значно більшій кількості функціоналу. Через те, що він є повністю програмним, його розробка не потребує витрат на різноманітні матеріали та обладнання, необхідні для виготовлення корпусу, схем, тощо.

Для впровадження ринкової реалізації проекту слід обрати альтернативу, яка передбачає розробку програмного продукту за допомогою нейронних мереж, а потім якісну рекламу та PR, сконцентровану навколо позитивних характеристиках даного програмного продукту, таких як низька ціна, більш істотний ефект покращення якості видачі, кросплатформеність і т.д.

З огляду на проведений аналіз, можна чітко сказати, що подальша імплементація проекту є доцільною, адже він може знайти свою цільову аудиторію та зайняти місце на ринку.

ВИСНОВКИ

Під час написання даної магістерської дисертації було розглянуто алгоритми колаборативної фільтрації, а саме: Item-Based, User-Baser, NMF, PMF, BPMF, Reg SVD та Slope One. Було надано короткий опис кожного з перерахованих алгоритмів.

Також, окремим розділом, було коротко розглянуто рекомендаційні системи.

Тепер давайте перейдемо до розгляду програмного продукту(Рисунок 3.1), розробленого в ході виконання даної роботи.

В ході роботи над даною магістерською дисертацією було розроблено програмний продукт, що включає в себе реалізацію набору алгоритмів колаборативної фільтрації.

Також, в межах написання роботи, на основі набору даних, наданого компанією Netflix для конкурсу NetflixPrize та набору даних сайту MovieLand було згенеровано тестові набори даних для тестування алгоритмів колаборативної фільтрації у різних ситуаціях, таких як:

- різна кількість користувачів (від 1 000 до 10 000)
- різна кількість фільмів (від 500 до 5 000)
- різною щільністю набору даних (від 0.5% до 5%)
- комбінації попереднії пунктів

Було проведено оцінку точності алгоритму та побудовано графіки залежності алгоритмів від кількості користувачів(Рисунок 3.2), кількості фільмів(Рисунок 3.3) та щільності(Рисунок 3.4) набору даних.

Графічно показано (Рисунок 3.5) залежність MAE (середня абсолютна похибка) від кількості користувачів, елементів та щільності. Також, до графічних даних наведено висновки.

Наведено графіки залежності швидкості роботи алгоритмів (Рисунок 3.6 – 3.12) від кількості користувачів, елементів та щільності набору даних.

Графічно показано ситуації, в яких той чи інший алгоритм працює найкраще (Рисунок 3.13 – 3.14) в залежності від обмежень у часі (1 хвилина та 5 хвилин).

Представлено висновки щодо точності та швидкодії алгоритмів, базуючись на експериментах.

Сформовано таблицю з характеристиками досліджуваних алгоритмів (Таблиця 5.1), базуючись на проведених дослідження. З цієї таблиці видно, що алгоритми Reg SVB, PMF та BPMF маюсь високу залежність від щільності та розмірів наборів даних і повільно працюють, але дають дуже гарні результати на розріджений наборах даних. NMF значно меншу залежність від розмірів та щільності набору даних, та має кращу точність на щільних наборах даних ніж попередня група алгоритмів, але гіршу точність на розріджений наборах (які, зазвичай, трапляються частіше). Slope One має скромніші результати з точності, особливо на щільних наборах даних, але він потребує найменше часу з усіх перерахованих алгоритмів на обчислення, та витрачає найменше пам’яті, що робить його дуже зручним у використанні, коли потрібно отримати якісний результат у найкоротші терміни.

Розглянуто можливість комбінації методів колаборативної фільтрації для мінімізації проблем, з якими зустрічаються окремі алгоритми.

Наведено розділ з аналізом ринку та підготовкою стратегії для створення стартап-проекту.

Новизною роботи є те, що порівняно такий набір алгоритмів, які до цього ще не порівнювались в одній роботі за такої кількості проведених експериментів на наборів даних.

Отже, було проведено дослідження алгоритмів колаборативної фільтрації, проведено аналіз їх точності та швидкодії в залежності від різних параметрів.

ПЕРЕЛІК ПОСИЛАНЬ

1. Рекомендаційна система — Вікіпедія [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/%D0%A0%D0%B5%D0%BA%D0%BE%D0%BC%D0%B5%D0%BD%D0%B4%D0%B0%D1%86%D1%96%D0%B9%D0%BD%D0%B0_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0. – Назва з екрану.
2. Netflix Prize [Електронний ресурс]. – Режим доступу: http://www.netflixprize.com/. – Назва з екрану.
3. Колаборативна фільтрація — Вікіпедія [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BB%D0%B0%D0%B1%D0%BE%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%B0_%D1%84%D1%96%D0%BB%D1%8C%D1%82%D1%80%D0%B0%D1%86%D1%96%D1%8F.– Назва з екрану.
4. Recommender Systems — User-Based and Item-Based Collaborative Filtering [Електронний ресурс]. – Режим доступу: https://medium.com/@cfpinela/recommender-systems-user-based-and-item-based-collaborative-filtering-5d5f375a127f – Назва з екрану
5. Как работают рекомендательные системы. Лекция в Яндексе / Блог компании Яндекс / Хабр [Електронний ресурс]. – Режим доступу: https://habr.com/company/yandex/blog/241455/– Назва з екрану
6. R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In Advances in Neural Information Processing Systems, 2008.
7. Singular Value decomposition (SVD) in recommender systems for Non-math-statistics-programming. [Електронний ресурс]. – Режим доступу:

https://medium.com/@m_n_malaeb/singular-value-decomposition-svd-in-recommender-systems-for-non-math-statistics-programming-4a622de653e9 - Назва з екрану

8. D. Lemire and A. Maclachlan. Slope one predictors for online rating-based collaborative
fitering. Society for Industrial Mathematics, 05:471 – 480, 2005.
9. D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems. MIT Press, 2001.
10. Управління проектами [Електронний ресурс]. – Режим доступу: http://uk.wikipedia.org/wiki/Управління_проектами. – Назва з екрану.
11. Діаграма Ганта [Електронний ресурс]. – Режим доступу: http://uk.wikipedia.org/wiki/Діаграма_Ганта. – Назва з екрану.
12. Прасолов А.П.. Методи колаборативної фільтрації у рекомендаційних системах / А.П.Прасолов. // Міжнародний науковий журнал "Інтернаука". – 2018. – №8.
13. Mean absolute error – Wikipedia [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Mean_absolute_error. – Назва з екрану.
14. Root-mean-square deviation – Wikipedia [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Root-mean-square_deviation. – Назва з екрану.
15. Z. Huang, D. Zeng, and H. Chen. A comparison of collaborative-_ltering recommendation algorithms for e-commerce. IEEE Intelligent Systems, 22:68-78, 2007.
16. N. D. Lawrence and R. Urtasun. Non-linear matrix factorization with gaussian processes.In Proceedings of the 26th Annual International Conference on Machine Learning, 2009.
17. Hiroshi Shimodaira, Similarity and recommender systems School ofInformatics,The University of Edinburgh, 2014.

18. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms”, Tenth International World Wide Web Conference (WWW10), May 1-5, 2001, Hong Kong.
19. K. Miyahara and M. J. Pazzani. Improvement of collaborative _ltering with the simple bayesian classi_er 1. (11), 2002.
20. X. Su and T. M. Khoshgoftaar. A survey of collaborative _ltering techniques. Adv. In Artif. Intell., 2009:4:2{4:2, January 2009.